SLIDE
Research Methodology in I.T.
Lecture 04 – Introduction to Text Reuse and Plagiarism Detection

Authors: Ms. Iqra Muneer and Dr. Rao Muhammad Adeel Nawab
Instructor: Dr. Rao Muhammad Adeel Nawab

SLIDE
Lecture Outline

- **Basics of Text Reuse**
- **Basics of Plagiarism**
- **Data Annotation for Text Reuse Detection**
- **Methods for Text Reuse and Plagiarism Detection**
- **Evaluation Measures**
- **Treating the Problem of Text Reuse / Plagiarism Detection as Machine Learning Problem – A Step by Step Example**

SLIDE
==========
Basics of Text Reuse
==========

SLIDE
Text Reuse - Definition

- **The process of creating a new text (or document) using the existing one(s)**
  - **Original Text (or Source Text)**
    - **The text which is used to create new text**
  - **Derived Text (or Reused Text)**
    - **The text created by reusing the original text(s)**
  - **Reuse can also be of**
    - **Idea**
    - **Image**
    - **Concept**
    - **Movies**
    - **Features etc.**

SLIDE
Text Reuse - Acceptable vs Non-Acceptable

- **Journalism**
  - **Text reuse is a common practice**
  - **Newspapers use text(s) provided by News Agencies to write newspaper articles**
- **Plagiarism**
  - **Unacknowledged text reuse is not acceptable**

SLIDE
Text Reuse in Journalism

- **News Agency**
  - **An organization that collects news items and distributes them to newspapers or broadcasters**
- **Text Reuse in Journalism**
  - **Newspapers use articles provided by News Agencies to write newspaper stories (or news articles)**
  - **Text reuse is a common and legitimate practice in the domain of Journalism**

SLIDE
Two Levels of Rewrite in Journalism

- **Derived vs Non-Derived**
  - **Derived**
    - **The Newspaper story was created by barrowing the text(s) from News Agencies**
    - **Non-Derived**
      - **The Newspaper story is written independently and doesn't barrow any text from News Agencies**

SLIDE
Three Levels of Rewrite in Journalism

- **Derived Category can be further divided into**
  - **Wholly Derived**

        - **News Agency text is the only source for the reused Newspaper text, which means it is a verbatim (or exact) copy of the News Agency text**
        - **In this case, most of the reused text is word-to-word copy of the source text**
    - **Partially Derived**
        - **The Newspaper text has been either derived from more than one News Agency or most of the text is paraphrased by the editor when rewriting from News Agency text source**
- **Non-Derived**
        - **The News Agency text has not been used in the production of the Newspaper text (though words may still co-occur in both documents), it has completely different facts and figures or is heavily paraphrased from the News Agency's copy**

SLIDE
Text Reuse – Importance

- **Large digital repositories are readily available, making it easier to text reuse and hard to detect it**
- **Powerful text editors are making it easier to rewrite / modify text**
- **Freely available Machine Translation systems are helping people to easily even reuse text written in language that they don't know**
- **Automatic text altering tools are making it easier to quickly modify text for reuse**

SLIDE
Text Reuse - Applications

- **Plagiarism Detection**
    - **Detecting unacknowledged reuse of text particularly in academia**
- **Duplicate (or Near-duplicate) Document Detection**
    - **For example, removing duplicate or near-duplicate documents from the set of documents returned by a Search Engine (or Information Retrieval System) against a user query**
- **Copyright infringement detection**

SLIDE
Text Reuse Detection - Task

- **Given**
    - **A text pair, Text 1 and Text 2 (input)**
- **Find**

- **how much text has been reused from Original (Text 1) to create Text 2 (output) i.e. goal is to identify the level of text reuse**

SLIDE
Text Reuse – Input and Output

- **Input**
  - **Text Pair (Text 1 and Text 2)**
- **Output**
  - **For two levels of text reuse**
    - **Derived / Non-Derived**
  - **For three levels of text reuse**
    - **Wholly Derived / Partially Derived / Non-Derived**
- **Goal**
  - **Identify the level of text reuse**

SLIDE
Text Reuse - Granularity

- **Text reuse may occur at five levels**
  1. **Word level**
  2. **Phrasal level**
  3. **Sentence level**
  4. **Passage / Paragraph level**
  5. **Document level**

SLIDE
Example 01 – Text Reuse at Word level

- **Input**
  - **Text 1**
    - **Meal**
  - **Text 2**
    - **Food**
- **Output**
  - **Derived**

SLIDE
Example 02 – Text Reuse at Word level

- **Input**
  - **Text 1**
    - **Meal**
  - **Text 2**
    - **Butter**
- **Output**
  - **Non-Derived**

SLIDE
Example 03 – Text Reuse at Word level

- **Input**
  - **Text 1**
    - **Like**
  - **Text 2**
    - **Nice**
- **Output**
  - **Derived**

SLIDE
Example 04 – Text Reuse at Word level

- **Input**
  - **Text 1**
    - **people**
  - **Text 2**
    - **Audience**
- **Output**
  - **Derived**

SLIDE
Example 05 – Text Reuse at Word level

- **Input**
  - **Text 1**
    - **Dinner**
  - **Text 2**
    - **Gathering**
- **Output**
  - **Non-Derived**

SLIDE
Example 06 – Text Reuse at Word level

- **Input**
  - **Text 1**
    - **Dinner**
  - **Text 2**
    - **Food**
- **Output**
  - **Derived**

SLIDE
Example 01 – Text Reuse at Phrasal level

- **Input**
  - **Text 1**
    - **A story as old as time.**

- **Text 2**
  - **A tale as old as time.**
- **Output**
  - **Derived**

SLIDE
Example 02 – Text Reuse at Phrasal level

- **Input**
  - **Text 1**
    - **A story as old as time.**
  - **Text 2**
    - **This is the story of old times.**
- **Output**
  - **Non Derived**

SLIDE
Example 03 – Text Reuse at Phrasal level

- **Input**
  - **Text 1**
    - **Reading a book.**
  - **Text 2**
    - **Reding an article.**
- **Output**
  - **Non Derived**

SLIDE
Example 04 – Text Reuse at Phrasal level

- **Input**
  - **Text 1**
    - **Ambling in the rain**
  - **Text 2**
    - **Walking in the rain**
- **Output**
  - **Derived**

SLIDE
Example 05 – Text Reuse at Phrasal level

- **Input**
  - **Text 1**
    - **The love of my life**
  - **Text 2**
    - **The crush of my life**
- **Output**
  - **Derived**

SLIDE
Example 06 – Text Reuse at Phrasal level

- **Input**
  - **Text 1**
    - **The love of my life**
  - **Text 2**
    - **The journey of love is very long**
- **Output**
  - **Non Derived**

SLIDE
Example 01 – Text Reuse at Sentence level

- **Input**
  - **Text 1**
    - **What is your age?**
  - **Text 2**
    - **How old are you?**
- **Output**
  - **Derived**

SLIDE
Example 02 – Text Reuse at Sentence level

- **Input**
  - **Text 1**
    - **Will it snow tomorrow?**
  - **Text 2**
    - **The weather prediction is quiet storming, what do you think about snow in the upcoming days?**
- **Output**
  - **Non Derived**

SLIDE
Example 03 – Text Reuse at Sentence level

- **Input**
  - **Text 1**
    - **Your car is nice**
  - **Text 2**
    - **I like your car**
- **Output**
  - **Derived**

SLIDE
Example 04 – Text Reuse at Sentence level

- **Input**
  - **Text 1**
    - **He said that sit-ins have caused a huge loss to national economy and the nation is depressed**
  - **Text 2**
    - **Prime minister said "sit-ins have caused a huge loss to national economy and the nation is depressed."**

- **Output**
  - **Derived**

SLIDE
Example 05 – Text Reuse at Sentence level

- **Input**
  - **Text 1**
    - **He said that sit-ins have caused a huge loss to national economy and the nation is depressed**
  - **Text 2**
    - **Prime minister said "sit-ins have caused a huge loss to national economy and the nation is depressed."**
- **Output**
  - **Derived**

SLIDE
Example 06 – Text Reuse at Sentence level

- **Input**
  - **Text 1**
    - **Baluchistan successfully holds 3rd round of LG elections**
  - **Text 2**
    - **Plots in the municipal elections, the PML-N won the third stage**
- **Output**
  - **Non Derived**

SLIDE
Example 07 – Text Reuse at Sentence level

- **Input**
  - **Text 1**
    - **His body was handed over to the heirs after legal formalities**
  - **Text 2**

        - **The body was handed over to the heirs.**
- **Output**
    - **Derived**

SLIDE
Example 08 – Text Reuse at Sentence level

- **Input**
    - **Text 1**
        - **NAB Chairman determined to root out corruption from society**
    - **Text 2**
        - **Honest people will send representatives in Parliament will help eliminate corruption: NAB.**
- **Output**
    - **Non Derived**

SLIDE
Example 01 – Text Reuse at Passage level

- **Input**
    - **Text 1**
        - **Cognizant of the need to accord greater attention towards protection of the vulnerable and marginalized segments of society, the government is committed to make every possible effort to put in place effective legal, economic and social frameworks so as to ensure protection of human rights," he said in his message on the occasion of International Human Rights Day (December 10)**
    - **Text 2**
        - **On the occasion of Human Rights Day, the Constitution of Pakistan, he said in his message to the citizens based on race, color or race, regardless of the guarantees. To ensure the protection of human rights as possible to provide effective legal, economic and social framework is determined. He is the celebration of Human Rights Day, on every level, promotion and protection of human rights is our solid commitment.**
- **Output**
    - **Derived**

SLIDE

Example 02 – Text Reuse at Passage level

- **Input**
  - **Text 1**
    - **Addressing a ceremony held in honour of the winners of Nobel Peace Prize-2014 and televised from Oslo, Norway, she expressed the resolve to continue her struggle for bringing all girls and boys in the education net and to fight for their rights.**
  - **Text 2**
    - **Pakistan and India live in peace, I am sure that we stop the progress. But if we do not succeed one another so that no country would be able to move ahead, both the issues of poverty, lack of education of children and women are denied basic rights, and we have these problems together to solve**
- **Output**
  - **Non Derived**

SLIDE

Example 03 – Text Reuse at Passage level

- **Input**
  - **Text 1**
    - **He said he had held very good meetings and talks with Iranian Minister of Economy and Finance Ali Tayebnia during his visit to Pakistan. He said that there were agreements with Chinese CNPC oil company recently for laying 700 kms gas pipeline from Gwadar port, 70 kms from Iran border, and Nawabshah. He said Chinese company envoy arrived on Tuesday to arrange preliminaries for the project, announcing that the company will complete the 700km pipeline in 24 months**
  - **Text 2**
    - **LNG terminal at the port in the first phase while the second phase will be held from Gwadar to Nawabshah 700 kilometers of 42-inch diameter pipeline will be laid. He said that because of international sanctions imposed on Iran Pakistan to fulfill its part of the project completed Despite the government's efforts could relate to this project Bank, International Contractors and Equipment Suppliers are not willing to work. Is expected to work on the project would be started soon.**

- **Output**
  - **Derived**

SLIDE
Example 04 – Text Reuse at Passage level

- **Input**
  - **Text 1**
    - **We are contacting the PTI leadership for starting the dialogue," he added. Welcoming the PTI's decision to resume talks, Dar said there was no impediment in this regard as PTI chief Imran Khan had backed out from the unconstitutional and illegal demand for the resignation of Prime Minister or for going on a one-month leave.**
  - **Text 2**
    - **'Movement and the government of rigging the kumbynh judicial commission investigating judge who will not be recognized, ISI or IB representatives of the Commission only the Commission can add your own, we would not demand anything, "Imran Mohib The evidence of Pakistani and sit-ins and rallies to protest the talks ended. If they do not agree then the negotiations' unconstitutional demand the resignation of the Minister of Justice to withdraw any unconstitutional is welcome, we will not discuss the talks and hope that justice will not demand any unconstitutional.**
- **Output**
  - **Non Derived**

SLIDE
Example 05– Text Reuse at Passage level

- **Input**
  - **Text 1**
    - **The participants, Malala said she was proud to be the youngest-ever Nobel Peace Prize recipient. She thanked her parents for providing all kind of support in getting education, saying "I thank them for not clipping my wings and letting me fly." Stressing the need to make joint efforts for imparting quality education to girls and boys without fear and discrimination, Malala vowed to work for protection of children's rights not only in Pakistan but across the world with more vigour and dedication.**

- **Text 2**
  - **After completing his studies at the prime minister's wish, to promote education seriously consider a global leader. The international community to pay special attention to education. Why can not give arms to provide easy and scripture. Why powerful countries are weak in peace. We are living in contemporary art, nothing is impossible, and I thank all my fans. I am grateful to my teachers and parents who gave me a chance to excel and education. This is a very happy day for me, I'm the youngest Pakistani and Pashtun girl who won it. Kailash Satyarthi champions are fighting for the rights of children. I am glad that we can work together.**

- **Output**
  - **Derived**

SLIDE

Example 06– Text Reuse at Passage level

- **Input**
  - **Text 1**
    - **According to senior police officials, the reason behind the murder has not been ascertained yet. Several police teams led by senior police officials were conducting raids at various places for the arrest of killers. Meanwhile, MQM has lodged protest in Sialkot city against the murder and blocked traffic on various roads of Sialkot. They were demanding early arrest of the killers.**
  - **Text 2**
    - **Praltaf Anwar Hussain said the BOA was the senior partner and sincere testimony of their organization is a senior fellow lost. The tortured bodies of our workers have poured into the streets of the patient is recommended. Altaf Hussain has warned that if not stopped killing our workers in the province of Punjab, including the prime minister will not enter into any Minister Sindh**
- **Output**
  - **Non Derived**

SLIDE

Example 01– Text Reuse at Document level

- **Input**

- **Document 1**
  - **She expressed the resolve to continue her struggle for bringing all girls and boys in the education net and to fight for their rights. She also recalled her struggle for getting education in Swat and Taliban's terror in the valley, who threatened girls to stop getting education.**
- **Document 2**
  - **But, she decided to stand up against them and succeeded, she added. "Terrorists failed in their nefarious designs," Malala said adding that was not alone as she was the voice of 66 million girls. Taliban, she said, blew up schools in Swat with bombs and rockets and they misused the name of Islam which was a religion peace, tolerance, brotherhood and humanity, urging the followers to get knowledge, education and discover new Praltaf Anwar Hussain said the BOA was the senior partner and sincere testimony of their organization is a senior fellow lost. The tortured bodies of our workers have poured into the streets of the patient is recommended. Altaf Hussain has warned that if not stopped killing our workers in the province of Punjab, including the prime minister will not enter into any Minister Sindh.**

- **Output**
  - **Derived**

SLIDE

Example 02– Text Reuse at Document level

- **Input**
  - **Document 1**
    - **Chairman Norwegian Nobel Peace Committee Thirdborn Jagland awarded the winners with gold medals and prizes in a widely televised-ceremony from Oslo, Norway. He highlighted efforts of Malala and Kailash for protecting children's rights and bringing all girls and boys in the education net. He said Malala faced Taliban in Swat, who were threatening to keep her away from education and even made an attempt on her life. She, however exhibited great courage and continued studies, besides advocating for girls' education.**

- **Document 2**
  - **It is time that education should take place, then do not raise any action against education. I want peace in every corner of the world, education is a key component of basic life henna on their hands, the formula used to calculate. I want that women be given equal rights, the award is for frightened children who want peace. Our Prophet Mohammad is the messenger of peace, I decided to speak out against the Taliban, and hundreds of schools were destroyed by militants in Swat, once a tourist paradise of Swat was killed by terrorists. Girls' education was stopped in Swat, militants tried to stop us, me and my friends were attacked, our voice has been compared to the Taliban, the Taliban's ideology not only won their shots prevail so, this story is not just me so many other girls, deprived of education stand to hear children's voices, this time will not be afraid and do virtually anything. Swat was always eager to learn and inventions. It is time that education should take place, then do not raise any action against education. I want peace in every corner of the world, education is a key component of basic life henna on their hands, the formula used to calculate. I want that women be given equal rights, the award is for frightened children who want peace. Our Prophet Mohammad is the messenger of peace, I decided to speak out against the Taliban, and hundreds of schools were destroyed by militants in Swat, once a tourist paradise of Swat was killed by terrorists.**

- **Output**
  - **Non Derived**

SLIDE

Example 03– Text Reuse at Document level

- **Input**
  - **Document 1**
    - **Around 500 family members of victims of Indian state repression along with human rights activists and members of Dal Khalsa during a march in Amritsar said that human rights abuses committed in Punjab and Kashmir were not random but were carried out as**

a matter of Indian state policy. They said that they would approach Barack Obama during his forthcoming visit to India on January 26, next year,KMS reported.

- **Document 2**
  - **The Indian state of Kashmir my dyasrus about five hundred families of victims of terrorism on human rights activists and members of Dal Khalsa accompanied the rally in Amritsar Punjab and Kashmir protesters said regular human rights violations Indian state policy being.**

- **Output**
  - **Derived**

SLIDE

Example 04– Text Reuse at Document level

- **Input**
  - **Document 1**
    - **In the decision it was stated that the counsel for the petitioner was confronted with the maintainability of these petitions in the light of the restrictions contained in article 225 of the Constitution on throwing a challenge to the election results other than by way of election petition before the Election Tribunal and further whether the results of the entire general elections for the national and the provincial assemblies could be annulled under any provision of the Constitution or law.**
  - **Document 2**
    - **The decision that the applicants lawyer has no way to hear election petitions, but tribunals do not give satisfactory answers to the question whether the context of Article 225 and whether national and annulled the results of the Assembly can be given.**
- **Output**
  - **Non Derived**

SLIDE

Text Reuse - Types

- **Local Text Reuse vs Global Text Reuse**
  - **Local Text Reuse**
    - **When amount of text reused is detected at sentence/passage level**

- **Global Text Reuse**
  - **When amount of text reused is detected at document level**

SLIDE
Text Reuse - Types

- **Mono-lingual Text Reuse vs Cross-lingual Text Reuse**
  - **Mono-lingual Text Reuse**
    - **When both the original and the reused text are in the same language**
  - **Cross-lingual Text Reuse**
    - **When the original text is in one language and the reused text is in another language**
    - **Cross-lingual text reuse can be carried out using**
      - **Automatic Translation (for e.g. Google Translator, BIgn Translator etc.)**
      - **Manual Translation**

SLIDE
Example - Mono-lingual Text Reuse

- **Input**
  - **Text 1**
    - **A dog bites a man**
  - **Text 2**
    - **A hound bites a person**
- **Output**
  - **Derived**
- **Note - Both source and reused texts are in the same language**

SLIDE
Example - Cross-lingual Text Reuse

- **Text 1**
  - **When source and reused texts are in different languages language**
    - **Source: A dog bites a man**
- **Text 2**
  - **ایک کتا ایک شخص کو کاٹتا ہے**
- **Output**
  - **Derived**
- **Note – Both source and reused texts are in different languages**

SLIDE

Example - Verbatim Copy/ Exact Copy (Mono-lingual Settings)

- **Text 1**
  - **He said that sit-ins have caused a huge loss to national economy and the nation is depressed**
- **Text 2**
  - **He said that sit-ins have caused a huge loss to national economy and the nation is depressed**

SLIDE

Example – Paraphrased Copy (Mono-lingual Settings)

- **Text 1**
  - **Prime Minister Nawaz Sharif recalled that in start of December he had announced reduction of 2.32 rupees per unit in prices of electricity, but the announcement was deferred because of Peshawar school tragedy**
- **Text 2**
  - **Prime Minister said that the price of electricity has been decreased by 2 rupees 32 paisas per unit. In the coming days [we] are trying to make electricity more affordable.**

SLIDE

Example - Independently Written (Mono-lingual Settings)

- **Text 1**
  - **Prime Minister Muhammad Nawaz Sharif Wednesday announced further reduction in the prices of petroleum products, which will be effective from January 1, 2015.**
- **Text 2**
  - **Prime Minister Nawaz Sharif has announced reduction in prices of petroleum products up to Rs. 14 per liter. Petrol 6.25, diesel 7.86, kerosene oil 11.26 rupees per liter cheaper [than before].**

SLIDE

Example - Verbatim Copy/ Exact Copy (Cross-lingual Settings)

- **Text 1**
  - **The chief minister said he would personally monitor the programmed of repair and construction of roads in rural areas and review the pace of progress on fortnightly basis.**
- **Text 2**

وزیراعلیٰ نے کہا کہ دیہی علاقوں میں سڑکوں کی تعمیر و مرمت کے پروگرام کی ذاتی
طور پر نگرانی کروں گا اور ہر پندرہ روز بعد پیشرفت کا جائزہ لوں گا۔

SLIDE
Example – Paraphrased Copy (Cross-lingual Settings)

- **Text 1**
  - **Prime Minister Nawaz Sharif recalled that in start of December he had announced reduction of 2.32 rupees per unit in prices of electricity, but the announcement was deferred because of Peshawar school tragedy.**
- **Text 2**

وزیراعظم نے کہا ہے کہ بجلی کی قیمت 2 روپے 32 پیسے فی یونٹ کم ہو چکی ہے۔
آئندہ دنوں میں بجلی مزید سستی کرنے کی کوشش کر رہے ہیں۔

SLIDE
Example - Independently Written (Cross-lingual Settings)

- **Text 1**
  - **Prime Minister Muhammad Nawaz Sharif Wednesday announced further reduction in the prices of petroleum products, which will be effective from January 1, 2015.**
- **Text 2**

وزیراعظم نواز شریف نے پٹرولیم مصنوعات کی قیمتوں میں 14 روپے فی لٹر تک کی
کا اعلان کر دیا۔ پٹرول 6.25، ڈیزل 7.86، مٹی کا تیل 11.26 روپے لٹر سستا۔

SLIDE
Your Turn

- **Write at least 12 examples for each of the following: (1) 6 examples for mono-lingual text reuse and (2) 6 examples for cross-lingual text reuse (2 Wholly Derived Examples, 2 Partially Derived Examples and 2 Non-Derived Examples)**
  - **Text Reuse at Word Level**
  - **Text Reuse at Phrasal Level**
  - **Text Reuse at Sentence Level**
  - **Text Reuse at Passage / Paragraph Level**
  - **Text Reuse at Document Level**

SLIDE
Summary – Basics of Text Reuse

- **Text Reuse is the process of creating a new text (or document) using the existing one(s) and it is reported to be on rise in recent years due to easy access to large online digital repositories**
- **Given a text pair (Text 1 and Text 2), Text 2 is said to be "Derived" from Text 1, if it is created using text from Text 1. On the other hand, Text 2 is said to be "Non-Derived" from Text 1, if it is interpedently written i.e. did not borrow text from Text 1**
- **Text Reuse may occur at five levels: (1) Word level, (2) Phrasal level, (3) Sentence level, (4) Passage / Paragraph level and (5) Document level**
- **Two main types of text reuse are: (1) Local Text Reuse - when amount of text reused is detected at sentence/passage level and (2) Global Text Reuse - when amount of text reused is detected at document level**
- **Text Reuse can be: (1) Mono-lingual Text Reuse - when both the original and the reused text are in the same language and (2) Cross-lingual Text Reuse - when the original text is in one language and the reused text is in another language**

SLIDE
==========
Basics of Plagiarism
==========

SLIDE
Plagiarism

- **Plagiarism is defined as the unacknowledged reuse of text**
- **Formal Definition**
  - **Copying another person's work exactly and presenting it as your own (without attributing it to the original author)**
- **Suspicious Document**
  - **The document suspected to contain plagiarism**
  - **Note that a suspicious document may or may not contain plagiarism**
- **Source Document(s)**
  - **The document(s) which were used to create the plagiarized document**

SLIDE
Plagiarism – Importance

- **In recent years, plagiarism has been reported to be on rise particularly in academia**
  - **Plagiarism detection systems are routinely used in universities to check students work for plagiarism**

SLIDE

Levels of Plagiarism

1. **Verbatim**
    a. **The original text is reused as verbatim (word to word copy) or with minor modifications to create the plagiarized document**
2. **Paraphrased Plagiarism**
    a. **The original text is heavily altered (or paraphrased) to create the plagiarized document**
    b. **Paraphrasing can be as**
        i. **Light Revision**
            1. **Source text is slightly paraphrased**
        ii. **Heavy Revision**
            1. **Source text is heavily paraphrased**
3. **Plagiarism of Idea**
    a. **The idea of the original text is reused without dependence on the words or form of the source**

SLIDE

Example - Verbatim Plagiarism

- **Text 1**
    - **In fact, of innumerable creatures predestined from the creation of the world to lay up a store of wealth for the British farmer, and a store of quite another sort for an immaculate Republican government.**
- **Text 2**
    - **Here lived innumerable creatures predestined from the creation of the world to lay up a store of wealth for the British farmer, and a store of quite another sort for an immaculate Republican government.**

SLIDE

Example - Paraphrased Plagiarism

- **Text 1**
    - **The number of foreign and domestic tourists in the Netherlands rose above 42 million in 2017, an increase of 9% and the sharpest growth rate since 2006, the national statistics office CBS reported on Wednesday" (DutchNews.nl, 2018).**

- **Text 2**

- **According to the national statistics office, the Netherlands experienced dramatic growth in tourist numbers in 2017. More than 42 million tourists travelled to or within the Netherlands that year, representing a 9% increase – the steepest in 12 years (DutchNews.nl, 2018).**

SLIDE
Example - Plagiarism of Idea

- **Text 1**
  - **The From a class perspective this put them [highwaymen] in an ambivalent position. In aspiring to that proud, if temporary, status of 'Gentleman of the Road', they did not question the inegalitarian hierarchy of their society. Yet their boldness of act and deed, in putting them outside the law as rebellious fugitives, revivified the 'animal spirits' of capitalism and became an essential part of the oppositional culture of working-class London, a serious obstacle to the formation of a tractable, obedient labour force. Therefore, it was not enough to hang them – the values they espoused or represented had to be challenged.**
- **Text 2**
  - **Peter Linebaugh argues that highwaymen represented a powerful challenge to the mores of capitalist society and inspired the rebelliousness of London's working class.**

SLIDE
Plagiarism Detection – Task

- **Given**
  - **A suspicious text (input)**
- **Identify**
  - **The source(s) of plagiarism**

SLIDE
Plagiarism Detection – Input and Output

- **Input**
  - **Suspicious Text**
- **Output**
  - **Plagiarized / Non-Plagiarized**

SLIDE
Plagiarism Detection - Two Levels of Rewrite

1. **Plagiarized**
   a. **When any type of plagiarism is occurred between documents they were called plagiarized**

2. **Non-Plagiarized**
    a. **When no type of plagiarism is occurred between documents they were called non plagiarized**

SLIDE

Plagiarism Detection - Four Levels of Rewrite

- **The Plagiarized cases can be further categorized into three categories**

1. **Near Copy**
    a. **When suspicious text is created by simply copying and pasting text from source document(s)**
2. **Light Revision**
    a. **When suspicious text is created by applying small modification like synonyms replacement and altering grammatical structure**
3. **Heavy Revision**
    a. **When suspicious text is created by rephrasing the text to generate the meaning**
        i. **It may include breaking source sentence into more than one sentences, margining two or more sentences into one, replacing words with appropriate synonyms or phrases, changing voice, changing tense etc.**
4. **Non-Plagiarized**
    a. **When suspicious text is written independently**

SLIDE

Important Note

- **Documents that independently written on the same topic are expected to have**
    - **Around 50% content overlap**

SLIDE

Example - Near Copy

- **Input**
    - **Text 1**
        - **A dog bites a man**
    - **Text 2**
        - **A dog bites a man**
    - **Output**
        - **Near Copy**

SLIDE

Example - Light Revision

- **Input**
  - **Text 1**
    - **A dog bites a man**
  - **Text 2**
    - **A hound bites by a person**
  - **Output**
    - **Light Revision**

SLIDE

Example - Heavy Revision

- **Input**
  - **Text 1**
    - **A dog bites a man**
  - **Text 2**
    - **A man was bitten by a dog**
  - **Output**
    - **Heavy Revision**

SLIDE

Example - Non-Plagiarized

- **Input**
  - **Text 1**
    - **A dog bites a man**
  - **Text 2**
    - **A man was injured while running on road followed by a dog**
  - **Output**
    - **Non-Plagiarized**

SLIDE

Type of Plagiarism Cases

- **There are three main types of plagiarism cases**
  - **Artificial**
    - **Artificial cases of plagiarism are generated by using Automatic Text Altering tools to obfuscate the source text for plagiarism**
    - **Three levels of rewrite**
      - **None Obfuscation**
        - **Automatic Text Altering tool simply copy and pastes text from source to create plagiarized document**
      - **Low Obfuscation**

        - **Automatic Text Altering tool lightly rephrases source text automatically before it is used to create plagiarized document**
    - **High Obfuscation**
        - **Automatic Text Altering tool heavily rephrases source text automatically before it is used to create plagiarized document**
- **Simulated / Manual**
    - **The original text is paraphrased by humans to create the cases of plagiarism**
- **Real**
    - **Real cases of plagiarism are those which occurred in the real world**
        - **For example, Karl-Theodor zu Guttenberg (German Defence Minister) PhD thesis proved plagiarized**
        - **URL:**https://www.theguardian.com/world/2011/mar/01/german-defence-minister-resigns-plagiarism

SLIDE

Example – Artificial (None Obfuscation)

- **Source**
    - **The first agrarian movement after the enactment of lex Licinia took place in the year 338, after the battle of Veseris in which the Latini and their allies were completely conquered.**
- **Suspicious**
    - **The first agrarian movement after the enactment of lex Licinia took place in the year 338, after the battle of Veseris in which the Latini and their allies were completely conquered.**

SLIDE

Example – Manual (Simulated Obfuscation)

- **Source**
    - **The emigrants who sailed with Gilbert were better fitted for a crusade than a colony, and, disappointed at not at once finding mines of gold and silver, many deserted; and soon there were not enough sailors to man all the four ships.**
- **Suspicious**
    - **The people who left their countries and sailed with Gilbert were more suited for fighting the crusades than for leading a settled life in the colonies. They were bitterly disappointed as it was not the America that they had expected. Since they did not immediately find gold and**

silver mines, many deserted. At one stage, there were not even enough man to help sail the four ships.

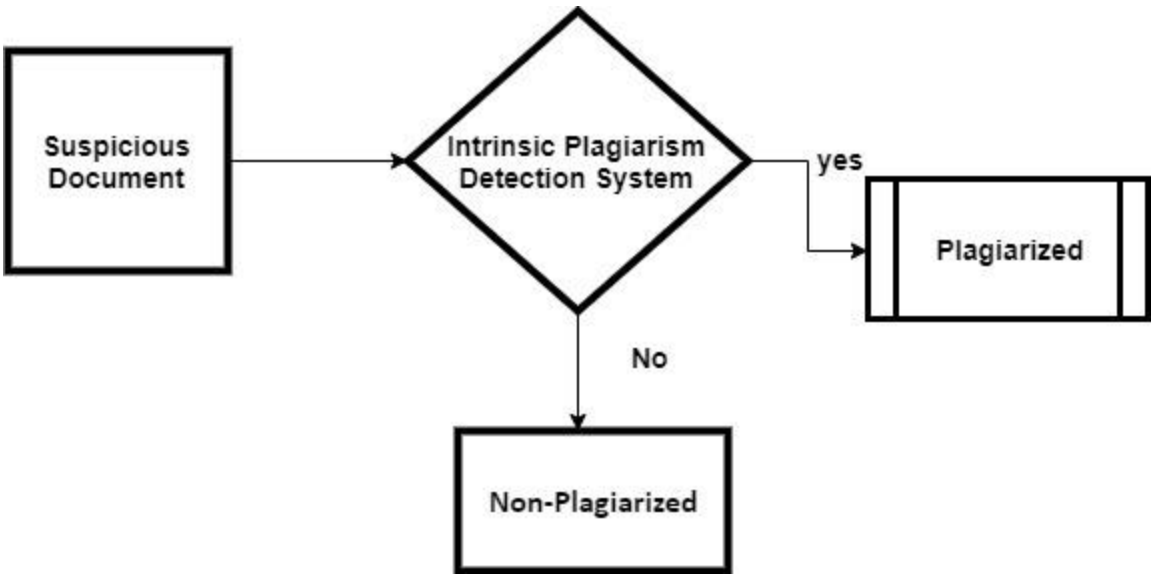
SLIDE
Real

- **Due to copyright issues it is impossible to have example of real case of plagiarism**

SLIDE
Types of Plagiarism Detection

- **Two main types of plagiarism detection**
  - **Intrinsic Plagiarism Detection**
    - **Checking that the entire document (or all the passages) were written by one single author**
    - **In case of intrinsic plagiarism detection, the focus is on identifying portion(s) of text whose writing style significantly differs from the remaining text in the suspicious document, which means that the entire document is not written by one single author and contains text written by other author(s).**
  - **Extrinsic Plagiarism Detection**
    - **Searching for the source(s) (or original text(s)) that were reused to create the suspicious document**
    - **Mainly involves comparison of the suspicious document with potential source documents**

SLIDE
Intrinsic Plagiarism Detection – Task



SLIDE
Intrinsic Plagiarism Detection – Task

- **Task**
  - **Given**
    - **A suspicious document (input)**
  - **Identify**
    - **Portion(s) of text whose writing style is significantly different from the remaining text (output)**

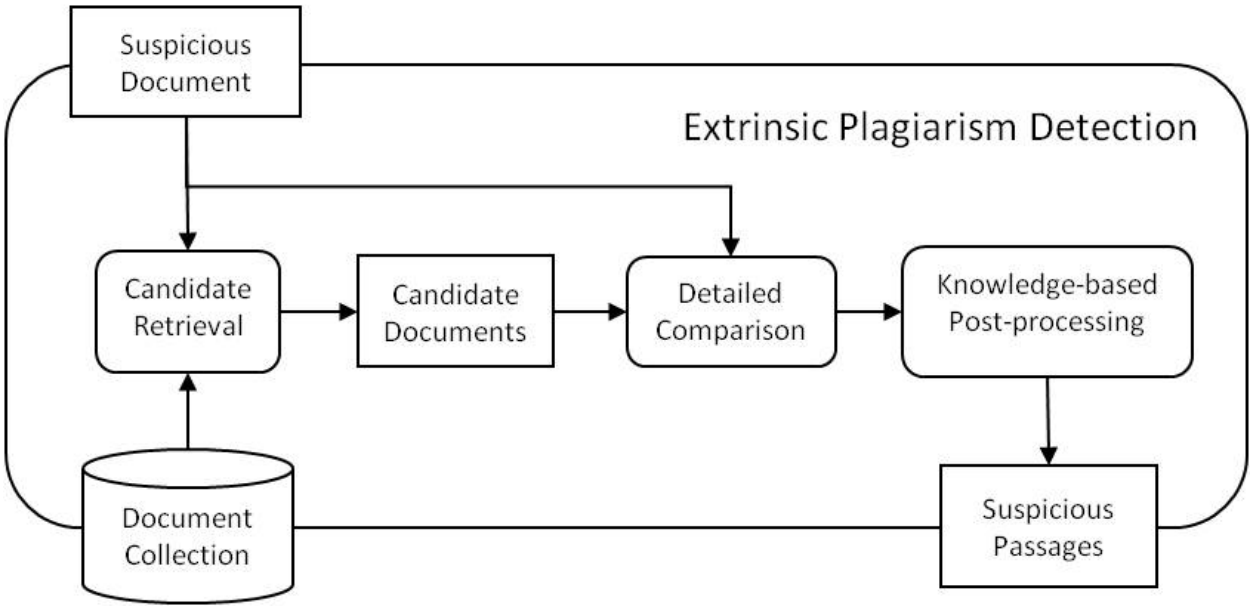
SLIDE
Intrinsic Plagiarism Detection – Input and Output

- **Input**
  - **A Suspicious Text**
- **Output**
  - **Portion(s) of text whose writing style is significantly different from the remaining text**
- **Note – If whose writing style is one or more portion(s) of text is significantly different from the remaining text then the suspicious document is plagiarized otherwise non-plagiarized**

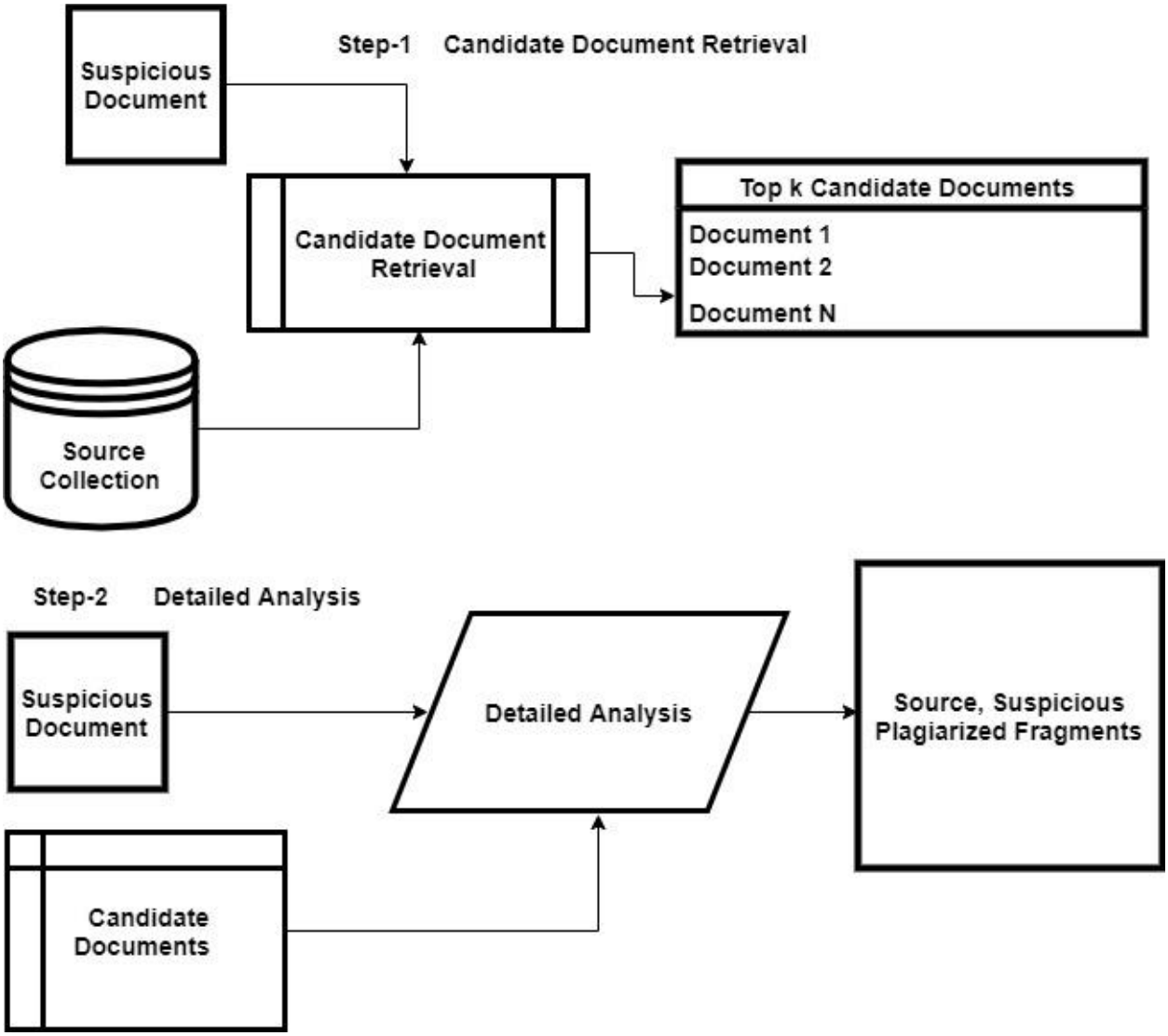
SLIDE
Example – Intrinsic Plagiarism Detection

- **Given (Suspicious Document)**
  - **Rasheed is my best friend. He lives in Lahore. He had got good education. He earned his PhD degree from one of the most prestigious, well reputed and renowned instructions of the world i.e. MIT, U.S.A. He is humble and nice. Rasheed always try to help others.**
- **Output**
  - **Suspicious Document is Plagiarized**
  - **Portion of text whose writing style is significantly different from remaining text**
    - **He earned his PhD degree from one of the most prestigious, well reputed and renowned instructions of the world i.e. MIT, U.S.A.**

SLIDE
Extrinsic Plagiarism Detection – Task



SLIDE
Extrinsic Plagiarism Detection – Task



SLIDE
Example – Extrinsic Plagiarism Detection

- **Given**
  - **Suspicious Collection**
    - **Containing two suspicious documents**
  - **Source Collection**
    - **Containing five source documents**

SLIDE

Example – Extrinsic Plagiarism Detection (Cont….)

- **Documents in Suspicious Collection**
  - **suspicious-document-01**
    - **Rashed is my best friend. He lives in Lahore. He had got good education. He earned his PhD degree from one of the most prestigious, well reputed and renowned instructions of the world i.e. MIT, U.S.A. He is humble and nice. Rashed always try to help others.**
  - **suspicious-document-02**
  - **Given a user query as input, Information Retrieval system aims to retrieve document(s) which are relevant to the user query to satisfy user's information need.**
- **Documents in Source Collection**
  - **source-document-01**
    - **Machine Learning a branch of AI. It is a hot research topic in the world.**
  - **source-document-02**
    - **MIT is one of the most prestigious, well reputed and renowned instructions of the world located in U.S.A.**
  - **source-document-03**
    - **My father name is Rao Nawab Akhtar. He is nice and humble. He always try to help others.**
  - **source-document-04**
    - **Natural Language Processing is a branch of AI. It has many potential applications in the real world.**
  - **source-document-05**
    - **Understanding is deeper than Love. A large number of people love you but only a few understand you.**

Example – Extrinsic Plagiarism Detection (Cont….)

- **Extrinsic Plagiarism Detection - Two Step Process**
  - **Candidate Document Retrieval**
    - **Aim**
      - **Identify potential source(s) of plagiarism**
    - **Potential Technique(s)**
      - **Use an Information Retrieval (IR) based approach**
  - **Detailed Analysis**
    - **Aim**
      - **Identify fragment(s) of source text(s) that were used to create the corresponding plagiarized fragments of text(s)**

- **Possible Technique(s)**
  - **Use a Pairwise Comparison Approach**

SLIDE

Example – Extrinsic Plagiarism Detection (Cont….)

- **Candidate Document Retrieval**
- **Goal**
  - **Identify top K source documents from the source collection which are potential sources of plagiarism**
    - **Here K = 3**
- **Candidate Document Retrieval System**
  - **Vector Space Model**
- **Given**
  - **Query**
    - **Two Separate Queries**
      - **suspicious-document-01**
      - **suspicious-document-02**
  - **Static Collection of Documents**
    - **Source Collection (containing 5 documents)**
- **Output of Candidate Document Retrieval System**
  - **suspicious-document-01 - Potential Candidate Source Document(s)**
    - **source-document-02**
    - **source-document-03**
    - **source-document-05**
  - **suspicious-document-01 - Potential Candidate Source Document(s)**
    - **source-document-01**
    - **source-document-04**
    - **source-document-05**

SLIDE

Example – Extrinsic Plagiarism Detection (Cont….)

- **Detailed Analysis**
- **Goal**
  - **Identify fragment(s) of source text(s) that were used to create the corresponding plagiarized fragments of text(s)**
- **Detailed Analysis Technique**
  - **Greedy String Tiling**
- **Given**
  - **Considering suspicious-document-01**
    - **Potential Candidate Source Documents**
      - **source-document-02**
      - **source-document-03**

            - **source-document-05**
    - **Considering suspicious-document-02**
        - **Potential Candidate Source Documents**
            - **source-document-01**
            - **source-document-04**
            - **source-document-05**
- **Output – Detailed Analysis**
- **Output – suspicious-document-01**
    - **suspicious-document-01 is Plagiarized**
    - **Source(s) of Plagiarism**
        - **source-document-02**
        - **source-document-03**
    - **Suspicious-Source Fragment Pairs**
        - **Fragment Pair 1 - suspicious-document-01, source-document-01**
            - **suspicious-document-01**
                - **one of the most prestigious, well reputed and renowned instructions of the world i.e. MIT, U.S.A**
            - **source-document-02**
                - **MIT is one of the most prestigious, well reputed and renowned instructions of the world located in U.S.A.**
        - **Fragment Pair 2 - suspicious-document-01, source-document-03**
            - **suspicious-document-01**
                - **He is humble and nice. Rashed always try to help others.**
            - **source-document-03**
                - **He is nice and humble. He always try to help others.**
- **Output – suspicious-document-02**
    - **suspicious-document-02 is Non-Plagiarized**

SLIDE
Shared Tasks on Text Reuse and Plagiarism
- **PAN**
    - **PAN is a series of scientific events and shared tasks on digital text forensics and stylometry**
    - **URL: https://pan.webis.de/**

SLIDE
Main Shared Tasks on Natural Language Processing
- **SemiEval**

- **URL: http://alt.qcri.org/semeval2019/**
- **CoNLL**
  - **URL: https://www.conll.org/2019**

SLIDE
Your Turn

- **Write at least 8 examples (at sentence level) for each of the following levels of rewrite in plagiarism**
  - **Near Copy**
  - **Light Revision**
  - **Heavy Revision**
  - **Non-Plagiarized**

SLIDE
Summary – Basics of Plagiarism

- **Plagiarism is defined as the unacknowledged reuse of text and in recent years it has been reported to be on rise. Consequently, plagiarism detection systems are routinely used by higher educational institutions to check students work for plagiarism**
- **Given a text pair (source text and suspicious text), suspicious text is said to be "Plagiarized" from source text, if it is created using text from source text. On the other hand, suspicious text is said to be "Non-Plagiarized" from source text, if it is interpedently written i.e. did not borrow text from source text**
- **There are three levels of Plagiarism: (1) Verbatim - the original text is reused as verbatim (word to word copy) or with minor modifications to create the plagiarized document, (2) Paraphrased Plagiarism - the original text is heavily altered (or paraphrased) to create the plagiarized document and (3) Plagiarism of Idea - the idea of the original text is reused without dependence on the words or form of the source**
- **There are three main types of Plagiarism Cases: (1) Artificial Cases of Plagiarism - are generated by using Automatic Text Altering tools to obfuscate the source text for plagiarism, (2) Simulated / Manual Cases of Plagiarism - the original text is paraphrased by humans to create the cases of plagiarism and (3) Real Cases of Plagiarms - are those which occurred in the real world**
- **Two main types of plagiarism detection are: (1) Intrinsic Plagiarism Detection - checking that the entire document (or all the passages) were written by one single author and (2) Extrinsic Plagiarism Detection - searching for the source(s) (or original text(s)) that were reused to create the suspicious document**

SLIDE
==========
Data Annotation for Text Reuse Detection
==========

SLIDE
Data vs Information

- **Data**
  - **Raw Facts and Figures**
  - **Varieties of Data**
    - **Structured Data**
      - **Data is stored, processed, and manipulated in a traditional Relational Database Management System (RDBMS)**
    - **Unstructured Data**
      - **Data that is commonly generated from human activities and doesn't fit into a structured database format**
    - **Semi-structured Data**
      - **Data doesn't fit into a structured database system, but is none-the-less structured by tags that are useful for creating a form of order and hierarchy in the data**
  - **Main Forms of Data**
    - **Text**
    - **Image**
    - **Video**
    - **Audio**
- **Information**
  - **Processed form of Data**

SLIDE
Data Annotation

- **Data annotation**
  - **a.k.a. data labeling / data tagging**
  - **is the process of labeling data to make it usable for machine learning?**
  - **is performed by domain experts (humans – a.k.a. annotators / taggers / raters)**
  - **requires a lot of effort, time and cost**
- **Example 01 – Data Annotation**
- **Raw Data**
  - **iPhone7 is a good mobile**
  - **Battery of this phone is bad**

- **I am using iphone7**
- **Data Annotation – Sentiment Analysis**
  - **iPhone7 is a good mobile, Positive**
  - **Battery of this phone is bad, Negative**
  - **I am using iphone7, Neutral**
- **Data Annotation – Gender Identification**
  - **iPhone7 is a good mobile, Male**
  - **Battery of this phone is bad, Male**
  - **I am using iphone7, Female**

SLIDE

Main Steps to Create Benchmark Annotated Dataset

- **Raw Data Collection**
  - **Data Source(s)**
  - **Cleaning of Data**
  - **Pre-processing of Data**
- **Annotation Process**
  - **Preparation of Annotation Guidelines**
  - **Annotations**
  - **Computing Inter-Annotator Agreement**
- **Corpus Standardization**

SLIDE

Example – Data Annotation for Text Reuse Detection

- **Raw Data Collection**
  - **Two Sources of Data**
    - **New Agencies articles**
      - **Associated Press of Pakistan (APP)**
      - **Independent News Agency (INP)**
    - **Newspapers stories**
      - **Express Newspaper**
      - **Jang Newspaper**
      - **Nawa e Waqt**
  - **Note – Multiple Newspapers can reuse text from one News Agency to produce their news stories**
  - **Raw Data Statistics**
    - **Collected 4 News Agency articles and 6 Newspaper stories**
  - **Cleaning of Raw Data**
    - **Remove HTML tags, hyperlinks, foreign language characters etc.**
  - **Pre-processing of Raw Data**
    - **No pre-processing was performed**

SLIDE
Example – Data Annotation for Text Reuse Detection

- **Below are cleaned and pre-processed documents (4 News Agency articles and 6 Newspaper stories)**

News Agency Article	Newspaper Story
A dog bites a man	A dog bites a person
A dog bites a man	A person was badly bitten by a dog
A dog bites a man	A person is badly injured while running at road followed by a hound, Independently written
I like your car	Your car is nice
This is my country	I live in Lahore, Pakistan
I like your car	I like your car

SLIDE
Example – Data Annotation for Text Reuse Detection
Annotation Process

- **Annotation Guidelines**
  - **Read text pair**
  - **Assign the label (from pre-defined set of labels) which maps to the "most dominating" label**
  - **Annotations**
    - **Standard practice in doing annotations is to have three annotators (A, B and C)**
      - **Can have more than three annotators**
    - **Characteristics of Annotators**
      - **Annotators must be domain experts**
      - **Annotators should be expert and / or native Speaker in the language in which text is written**
      - **Annotators should be of good qualification**
    - **Generally, annotations are carried out in three steps**
      - **Step 1: Annotators A and B annotate a subset of the dataset**
        - **discuss Conflicting and agreed text pairs and refine the annotation guidelines to reduce the conflicts**
      - **Step 2: Annotators A and B annotate the remaining dataset based on revised annotation guidelines**
      - **Step 3: Annotator C annotates conflicting documents**

SLIDE
Example – Data Annotation for Text Reuse Detection
Annotation Process

- **Annotations**
  - **Separately Give Text Pairs to Annotators A & B**
    - **Annotations by Annotator A**

News Agency Article	Newspaper Story	Annotator A
A dog bites a man	A dog bites a person	Wholly Derived
A dog bites a man	A person was badly bitten by a dog	Partially Derived
A dog bites a man	A person is badly injured while running at road followed by a hound, Independently written	Non - Derived
I like your car	Your car is nice	Non - Derived
This is my country	I live in Lahore, Pakistan	Partially Derived
I like your car	I like your car	Wholly Derived

SLIDE
Example – Data Annotation for Text Reuse Detection
Annotation Process

- **Annotations**
  - **Separately Give Text Pairs to Annotator A & B**
    - **Annotations by Annotator B**

News Agency Article	Newspaper Story	Annotator B
A dog bites a man	A dog bites a person	Wholly Derived
A dog bites a man	A person was badly bitten by a dog	Partially Derived
A dog bites a man	A person is badly injured while running at road followed by a hound, Independently written	Non - Derived
I like your car	Your car is nice	Partially Derived
This is my country	I live in Lahore, Pakistan	Non - Derived
I like your car	I like your car	Wholly Derived

SLIDE
Example – Data Annotation for Text Reuse Detection
Annotation Process

- **Inter-Annotator Agreement**
  - **Inter-Annotator Agreement (IAA) is a measure of how well two (or more) annotators can make the same annotation decision for a certain category**

$$IAA = \frac{count\ the\ number\ of\ ratings\ in\ agreement}{count\ the\ total\ number\ of\ ratings}$$

- **IAA is computed to derive two things**
  - **How easy was it to clearly delineate the category?**
  - **How trustworthy is the annotation?**

News Agency Article	Newspaper Story	Annotator A	Annotator B
A dog bites a man	A dog bites a person	Wholly Derived	Wholly Derived
A dog bites a man	A person was badly bitten by a dog	Partially Derived	Partially Derived
A dog bites a man	A person is badly injured while running at road followed by a hound, Independently written	Non - Derived	Non - Derived
I like your car	Your car is nice	Non - Derived	Partially Derived
This is my country	I live in Lahore, Pakistan	Partially Derived	Non - Derived
I like your car	I like your car	Wholly Derived	Wholly Derived

SLIDE
Example – Data Annotation for Text Reuse Detection
Annotation Process

  - **Combine Annotations of A & B to Compute Inter Annotator Agreement (IAA)**
- **Inter-Annotator Agreement =** $\frac{4}{6}$ **= 0.667**

SLIDE
Example – Data Annotation for Text Reuse Detection
Annotation Process

- **Conflict Resolution**
  - **Give only conflicted pairs to Annotator C**

News Agency Article	Newspaper Story	Annotator A	Annotator B	Annotator C
I like your car	Your car is nice	Partially Derived	Non Derived	Partially Derived
This is my country	I live in Lahore, Pakistan	Non Derived	Partially Derived	Non Derived

SLIDE
Example – Data Annotation for Text Reuse Detection
Annotation Process

- **Final Gold Standard Benchmark Corpus**

News Agency Article	Newspaper Story	Label
A dog bites a man	A dog bites a person	Wholly Derived
A dog bites a man	A person was badly bitten by a dog	Partially Derived
A dog bites a man	A person is badly injured while running at road followed by a hound, Independently written	Non Derived
I like your car	Your car is nice	Non Derived
This is my country	I live in Lahore, Pakistan	Partially Derived
I like your car	I like your car	Wholly Derived

SLIDE
Example – Corpus Standardization

- **Two Main Formats to Standardize Corpus**
  - **CSV**
  - **XML**

SLIDE
Example – Corpus Standardization

- **Corpus Standardization in CSV Format**

	A	B	C
1	Source	Derived	Label
2	A dog bite	A dog bites a person	Wholly Derived
3	A dog bite	A person was badly bitt	Partially Derived
4	A dog bite	A person is badly injure	Non-Derived
5	I like your	Your car is nice	Non-Derived
6	This is my	I live in Lahore, Pakistar	Partially Derived
7	I like your	I like your car	Wholly Derived
8	My favorit	My favorite subject is N	Wholly Derived
9	My favorit	I like NLP	Partially Derived
10	I like NLP	I am studying many cou	Non-Derived
11	Allama iqb	it was iqbal who awoke	Non-Derived
12	Balochista	Plots in the municipal e	Partially Derived
13	and he wa	World Cup has never re	Partially Derived
14	Syed Sulta	World Blind Cricket Cou	Wholly Derived
15	He said th	Raza Rabbani said the S	Wholly Derived
16	LHC-death	5 guilty of the death pe	Non-Derived
17			

SLIDE

Example – Corpus Standardization

- **Corpus Standardization in XML Format**

```xml
<?xml version="1.0" encoding="UTF-8"?>
<TRUE_document>
<r classification="WD" id="0001-1-1">
A dog bites a man</r><r classification="WD" id="0001P-1-1">
A dog bites a person</r>
<r classification="PD" id="0001-1-2">
A dog bites a man</r><r classification="PD" id="0001P-1-2">
A person was badly bitten by a dog</r>
<r classification="ND" id="0001-2-3">
A dog bites a man</r><r classification="ND" id="0001P-2-3">
A person is badly injured while running at road followed by a hound, Independently written</r>
<r classification="WD" id="0002-1-1">
I like your car</r><r classification="WD" id="0002P-1-1">
I like your car</r>
<r classification="PD" id="0002-1-2">
I like your car</r><r classification="PD" id="0002P-1-1">
your car is nice</r>
<r classification="ND" id="0002-1-1">
This is my country</r><r classification="ND" id="0002P-1-1">
I live in Lahore, Pakistan</r>
</TRUE_document>
```

SLIDE

Benchmark Corpora for Text Reuse and Plagiarism Detection

- **METER Corpus**
  - **Paper Title: The METER corpus: A corpus for analyzing journalistic text reuse**
- **SAC: Short Answer Corpus**

        - **Paper Title: Developing a corpus of plagiarized short answers**
    - **PAN Corpora**
        - **Can be downloaded from the PAN website: pan.webis.de**
    - **USTRC**
        - **Paper Title: USTRC: Urdu Short Text Reuse Corpus**

    - **CLEU: Cross lingual English Urdu Corpus**
        - **Paper Title: Design and Development of a Large Cross-Lingual Plagiarism Corpus for Urdu-English Language Pair**

SLIDE
Your Turn

- **Take a toy dataset of 12 examples and annotate them with four levels of rewrite by following the steps discussed in the lecture**
    - **Near Copy**
    - **Light Revision**
    - **Heavy Revision**
    - **Non-Plagiarized**

SLIDE
Summary – Data Annotation for Text Reuse Detection

- **Data Annotation**
    - **is the process of labeling data to make it usable for machine learning?**
    - **is performed by domain experts (humans – a.k.a. annotators / taggers / raters)**
    - **requires a lot of effort, time and cost**
- **Main Steps to Create Benchmark Annotated Dataset**
    - **Raw Data Collection**
        - **Data Source(s)**
        - **Cleaning of Data**
        - **Pre-processing of Data**
    - **Annotation Process**
        - **Preparation of Annotation Guidelines**
        - **Annotations**
        - **Computing Inter-Annotator Agreement**
    - **Corpus Standardization**

SLIDE
==========
Methods for Text Reuse and Plagiarism Detection
==========

SLIDE
Methods for Text Reuse and Plagiarism Detection

- **Given**
  - **A text pair (Text 1 and Text 2)**
- **Goal**
  - **Quantify the "degree of similarity" between text pair**
- **Note – High similarity score indicates that Text 2 was created using Text 1**

SLIDE
Example – N-gram Overlap Approach for Text Reuse and Plagiarism Detection

- **N-gram - Definition**
  - **An n-gram is a contiguous sequence of n items from a given sample of text**
  - **N-gram can be**
    - **Word based**
    - **Character based**
  - **N represents the length of N-gram**

SLIDE
Example – N-gram Generation from Input Text

- **Input Text**
  - **R u coming?**
- **Word Uni-grams (N = 1)**
  - **Tokenized Text:**
    - **R**
    - **u**
    - **coming**
    - **?**
  - **Set of Word Uni-grams = {R, u, coming, ?}**

SLIDE
Example – N-gram Generation from Input Text

- **Input Text**
  - **R u coming?**
- **Word Bi-grams (N = 2)**
  - **Tokenized Text:**

        - **R**
        - **u**
        - **coming**
        - **?**
    - **Set of Word Bi-grams = {R u, u coming, coming ?}**

SLIDE
Example – N-gram Generation from Input Text

- **Input Text**
    - **R u coming?**
- **Word Tri-grams (N = 3)**
    - **Tokenized Text:**
        - **R**
        - **u**
        - **coming**
        - **?**
    - **Set of Word Tri-grams = {R u coming, u coming ?}**

SLIDE
Example – N-gram Generation from Input Text

- **Input Text**
    - **R u coming?**
- **Character Tri-grams (N = 3)**
    - **Toeknized Text:**
    - **Note that space is also a character**
        - **R, ,u, ,c,o,m,I,n,g,?**
    - **Set of Character Tri-grams**
        - **{R u, u , co,com,omi,min,ing,gn?)**

SLIDE
Similarity Measures to Compute N-gram Overlap

- **A range of measures have been proposed including**
    - **Jaccard Similarity Co-efficient**
    - **Dice Similarity Co-efficient**
    - **Containment Similarity Co-efficient**
    - **Overlap Similarity Co-efficient**

SLIDE
Steps – Computing Similarity between Sets of N-grams using Overlap Similarity Co-efficient

- **Given**
    - **A Text Pair (Text 1 and Text 2)**

- **Steps - Commuting Similarity**
- **Step 1: Preprocess Input Text**
  - **Lower case**
  - **Punctuation Marks Removal**
- **Step 2: Convert Text 1 and Text 2 into Sets of N-grams**
- **Step 3: Compute Similarity Between Sets of N-grams using Overlap Similarity Co-efficient i.e. Quantify the Degree of Similarity**
  - **Overlap Similarity =** $= \frac{|S(\text{Text 1},n) \cap S(\text{Text 2},n)|}{\min\,(|S(Text\ 1,n)|,|S(\text{Text 2},n)|}$
- **Where S(Text 1, n) and S(Text 1, n) represent sets of n-grams of length n for Text 1 and Text 2 respectively.**

SLIDE

Example – Computing Similarity between Sets of N-grams using Overlap Similarity Co-efficient

- **Goal**
  - **Quantify the degree for similarity between text pair using N-gram Overlap Approach**
    - **Here**
      - **N-grams are word based**
      - **N = 1**
- **Given**
  - **A Text Pair**
    - **Text 1: A dog bites a man.**
    - **Text 2: A hound bites a man.**
- **Steps - Commuting Similarity**
- **Step 1: Preprocess Input Text**
    - **Lower case**
    - **Punctuation Marks Removal**
- **Text Pair Before Pre-processing**
    - **Text 1: A dog bites a man.**
    - **Text 2: A hound bites a man.**
- **Text Pair After Pre-processing**
    - **Text 1: a dog bites a man**
    - **Text 2: a hound bites a man**
- **Step 2: Convert Text 1 and Text 2 into Sets of N-grams**
    - **Text 1 Word Unigrams (S(Text 1, 1)) = {a, dog, bites, a, man}**
    - **Text 2 Word Unigrams (S(Text 2, 1)) = {a, hound, bites, a, man}**
- **Step 3: Compute Similarity Between Sets of N-grams using Overlap Similarity Co-efficient i.e. Quantify the Degree of Similarity**

- **Overlap Similarity =** $= \frac{|S(\text{Text 1},n) \cap S(\text{Text 2},n)|}{\min(|S(Text\ 1,n)|,|S(\text{Text 2},n)|}$
- **Overlap Similarity = |{a, dog, bites, a, man}| ∩|{a, hound, bites, a, man}| / min (|({a, dog, bites, a, man}|, (|{a, hound, bites, a, man}|)))**
- $\frac{|\{a,bites,a,man\}|}{\min(5,5)}$
- $\frac{4}{5}$ **= 0.80**

SLIDE

Example – Computing Similarity Between Sets of N-grams using Overlap Similarity Co-efficient

- **Goal**
  - **Quantify the degree for similarity between text pair using N-gram Overlap Approach**
    - **Here**
      - **N-grams are word based**
      - **N = 2**
- **Given**
  - **A Text Pair**
    - **Text 1: A dog bites a man.**
    - **Text 2: A hound bites a man.**
- **Steps - Commuting Similarity**
- **Step 1: Preprocess Input Text**
    - **Lower case**
    - **Punctuation Marks Removal**
- **Text Pair Before Pre-processing**
    - **Text 1: A dog bites a man.**
    - **Text 2: A hound bites a man.**
- **Text Pair After Pre-processing**
    - **Text 1: a dog bites a man**
    - **Text 2: a hound bites a man**
- **Step 2: Convert Text 1 and Text 2 into Sets of N-grams**
    - **Text 1 Word Unigrams (S(Text 1, 2)) = {a dog, dog bites, bites a, a man}**
    - **Text 2 Word Unigrams (S(Text 1, 2)) = {a hound , hound bites , bites a, a man}**
- **Step 3: Compute Similarity Between Sets of N-grams using Overlap Similarity Co-efficient i.e. Quantify the Degree of Similarity**
    - **Overlap Similarity =** $= \frac{|S(\text{Text 1},n) \cap S(\text{Text 2},n)|}{\min(|S(Text\ 1,n)|,|S(\text{Text 2},n)|}$
    - **Overlap Similarity = |{a dog, dog bites, bites a, a man}| ∩ |{a hound , hound bites , bites a, a man}| /**

min (|{a dog, dog bites, bites a, a man}|, |{a hound , hound bites , bites a, a man}|)))

- $\frac{|\{\text{bites a,a man}\}|}{\min(4,4)}$
- $\frac{2}{4}$ **= 0.5**

SLIDE

Example – Computing Similarity Between Sets of N-grams using Overlap Similarity Co-efficient

- **Goal**
  - **Quantify the degree for similarity between text pair using N-gram Overlap Approach**
    - **Here**
      - **N-grams are word based**
      - **N = 3**
- **Given**
  - **A Text Pair**
    - **Text 1: A dog bites a man.**
    - **Text 2: A hound bites a man.**
- **Steps - Commuting Similarity**
- **Step 1: Preprocess Input Text**
    - **Lower case**
    - **Punctuation Marks Removal**
- **Text Pair Before Pre-processing**
    - **Text 1: A dog bites a man.**
    - **Text 2: A hound bites a man.**
- **Text Pair After Pre-processing**
    - **Text 1: a dog bites a man**
    - **Text 2: a hound bites a man**
- **Step 2: Convert Text 1 and Text 2 into Sets of N-grams**
    - **Text 1 Word Trigrams (S(Text 1, 3)) = {a dog bites, dog bites a, bites a man}**
    - **Text 2 Word Trigrams (S(Text 1, 3)) = {a hound bites, hound bites a , bites a man}**
- **Step 3: Compute Similarity Between Sets of N-grams using Overlap Similarity Co-efficient i.e. Quantify the Degree of Similarity**
    - **Overlap Similarity =** $= \frac{|S(\text{Text 1},n) \cap S(\text{Text 2},n)|}{\min(|S(Text\ 1,n)|,|S(\text{Text 2},n)|}$
    - **Overlap Similarity = |{a dog bites, dog bites a, bites a man}| ∩ |{a hound bites, hound bites a , bites a man}|/ min (|{a dog bites, dog bites a, bites a man}|, |{a hound bites, hound bites a , bites a man}|)))**

- $\frac{|\{\text{bites a man}\}|}{\min(3,3)}$
- $\frac{1}{3} = 0.33$

SLIDE

Methods for Mono-lingual Text Reuse and Plagiarism Detection

- **Methods based on content**
  - **Word n-grams overlap**
  - **Vector Space Model**
  - **Methods based on string and sequence alignment**
  - **Longest common subsequence**
  - **Greedy String-Tiling**
  - **Global Alignment**
  - **Local Alignment**
- **Methods based on structure**
  - **Stop-words based n-grams overlap**
- **Methods based on style**
  - **Type token ratio**
  - **Token ratio**
  - **Sentence ratio**

SLIDE

Methods for Cross-lingual Text Reuse and Plagiarism Detection

- **Methods based on Syntax**
- **Cross-Language Character N-Gram**
- **Methods based on Dictionaries**
  - **Cross-Language Vector Space Method**
  - **Cross-Language Conceptual Thesaurus based Similarity**
  - **Cross-Language Knowledge Graph Analysis**
- **Methods based on Parallel Corpora**
  - **Cross-Language Alignment based Similarity Analysis**
  - **Cross-Language Latent Semantic Indexing**
  - **Cross-Language Kernel Canonical Correlation Analysis**
- **Methods based on Comparable Corpora**
  - **Cross-Language Explicit Semantic Analysis**
- **Methods based on Machine Translation**
  - **Translation + Monolingual Analysis**
- **Methods based on Word Embeddings**
  - **Cross-Language Conceptual Thesaurus based Similarity using Word Embeddings**
  - **Cross-Language Word Embeddings based Similarity**
  - **Cross-Language Word Embedding based Syntax Similarity**
- **Methods based on Deep Learning**

SLIDE
Your Turn

- **Considering the following example**
  - **Text 1: a dog bites a man**
  - **Text 2: a hound bites a man**
- **Using Longest Common Subsequence (LCS) Approach the "LCS" between two text pairs is:**
  - **a bites a man**
- **Computing Similarity**
  - **Similarity Score = len(LCS) / min (len(Text 1), len(Text 2))**
  - **Similarity Score = 4 / min (5, 5) = 4 / 5 = 0.80**
- **Your Task**
  - **is to take at least three text pairs and compute similarity score between then using the Longest Common Subsequence (LCS) Approach**

SLIDE
Summary – Methods for Text Reuse and Plagiarism Detection

- **Methods for Mono-lingual Text Reuse and Plagiarism Detection and be broadly categorized into: (1) Methods based on Content, (2) Methods based on Structure and (3) Methods based on Style**
- **Methods for Cross-lingual Text Reuse and Plagiarism Detection and be broadly categorized into: (1) Methods based on Syntax, (2) Cross-Language Character N-Grams, (3) Methods based on Dictionaries, (4) Methods based on Parallel Corpora, (5) Methods based on Comparable Corpora, (6) Methods based on Word Embeddings and (7) Methods based on Deep Learning**

SLIDE
==========
Evaluation Measures
==========

SLIDE
Precision

- **Precision (P) of a text reuse detection system is the proportion of the predicted positive cases that were correct.**
  - **P=**$\frac{TP}{TP+FP}$

SLIDE
Recall

- **Recall (R) of a text reuse detection system is defined as the proportion of positive cases that were correctly identified.**
    - **R =** $\frac{TP}{TP+FN}$

SLIDE
F_1 measure

- **F_1 measure is a specific relationship (harmonic mean) between precision (P) and recall (R).**
    - **$F_1 =$** $\frac{2*P*R}{P+R}$

SLIDE
Summary – Evaluation Measures

- **Evaluation of Text Reuse / Plagiarism Detection Systems is carried out using Precision, Recall and F_1 measures**
- **Note that in research papers (or thesis) mostly "weight average" Precision, Recall and F_1 scores are reported**

SLIDE
==========
Treating the Problem of Text Reuse / Plagiarism Detection as Machine Learning Problem – A Step by Step Example
==========

SLIDE
Treating the Problem of Text Reuse / Plagiarism Detection as Machine Learning Problem

- **Problem**
    - **Text Reuse Detection**
- **Input**
    - **A Text Pair**
- **Output**
    - **For Ternary Classification**
        - **Wholly Derived / Partially Derived / Non-Derived**
    - **For Binary Classification**
        - **Derived / Non-Derived**

SLIDE
Treating the Problem of Text Reuse Detection as Machine Learning Problem

- **Question?**
  - **How to Transform Text Reuse Detection Problem to Supervised Text Classification Task?**
- **For Supervised Text Classification Task**
  - **Output must be associated with the input i.e. dataset must be annotated**

SLIDE
Steps - Treating the Problem of Text Reuse Detection as Machine Learning Problem

- **Main Steps to Treat Text Reuse Detection as Supervised Text Classification Task**
- **Step 1: Pre-process input (i.e. text pair)**
- **Step 2: Feature Extraction from Input**
- **Step 3: Represent "extracted features" and "output / label" into a format that Machine Learning algorithms can understand (normally saved as CSV file)**
- **Step 4: Use "features" extracted in Stop 2 to train and test Machine Learning algorithms**

SLIDE
Experimental Setup

- **Problem of Text Reuse Detection is Treated as s Supervised Text Classification Task**
- **Two Versions of Classification**
  - **Binary Classification**
    - **Discriminate between two classes i.e. Derived vs Non-derived**
  - **Ternary Classification**
    - **Discriminate between three classes i.e. Wholly Derived / Partially Derived / Non-derived**
- **Dataset**
  - **File containing dataset in CSV format is called "data.csv"**
  - **15 instances (input + output)**
    - **Input – Text Pair**
    - **Output – Label**

Text 1	Text 2	Label
A dog bites a man	A dog bites a person	Wholly Derived
A dog bites a man	A person was badly bitten by a dog	Partially Derived
A dog bites a man	A person is badly injured while running at road followed by a hound, Independently written	Non Derived
I like your car	Your car is nice	Non Derived
This is my country	I live in Lahore, Pakistan	Partially Derived
I like your car	I like your car	Wholly Derived
My favorite subject is NLP	My favorite subject is NLP	Wholly Derived
My favorite subject is NLP	I like NLP	Partially Derived
I like NLP	I am studying many course but nlp an be top ranked	Non Derived
Allama iqbal is our national hero	it was iqbal who awoke the muslim with his poetry	Non Derived
Balochistan successfully holds 3rd round of LG elections	Plots in the municipal elections, the PML-N won the third stage	Partially Derived
and he was once looking forward to it," he reiterated.	World Cup has never refused: Shoaib Malik	Partially Derived
Syed Sultan Shah termed it a national tragedy.	World Blind Cricket Council chairman Syed Sultan Shah said that Peshawar is a national tragedy.	Wholly Derived
He said that increase in the tax is unconstitutional and violation of article 77.	Raza Rabbani said the Supreme Court's decision in the light of Article 77 is a violation of the decision.	Wholly Derived
LHC-death LHC dismisses appeals against death sentence	5 guilty of the death penalty rejected pleas for mercy Reporter Karachi to Islamabad? President's death convicted criminals 5 rejected pleas for mercy.	Non Derived

- **Technique (for Feature Extraction)**
  - **N-gram Overlap**
    - **N-grams are word based**
    - **N = 1 - 5**
- **Evaluation Measure**
  - **Precision**
  - **Recall**
  - **F_1**

SLIDE

Example - Treating the Problem of Text Reuse Detection as Machine Learning Problem

- **Main Steps to Treat Text Reuse Detection as Supervised Text Classification Task**
- **Step 1: Pre-process input (i.e. text pair)**

Text 1	Text 2
A dog bites a man	A dog bites a person
A dog bites a man	A person was badly bitten by a dog
A dog bites a man	A person is badly injured while running at road followed by a hound, Independently written
I like your car	Your car is nice
This is my country	I live in Lahore, Pakistan
I like your car	I like your car
My favorite subject is NLP	My favorite subject is NLP
My favorite subject is NLP	I like NLP
I like NLP	I am studying many course but nlp can be top ranked
Allama iqbal is our national hero	it was iqbal who awoke the muslim with his poetry
Balochistan successfully holds 3rd round of LG elections	Plots in the municipal elections, the PML-N won the third stage
and he was once looking forward to it," he reiterated.	World Cup has never refused: Shoaib Malik
Syed Sultan Shah termed it a national tragedy.	World Blind Cricket Council chairman Syed Sultan Shah said that Peshawar is a national tragedy.
He said that increase in the tax is unconstitutional and violation of article 77.	Raza Rabbani said the Supreme Court's decision in the light of Article 77 is a violation of the decision.
LHC-death LHC dismisses appeals against death sentence	5 guilty of the death penalty rejected pleas for mercy Reporter Karachi to Islamabad? President's death convicted criminals 5 rejected pleas for mercy.

- **Step 2: Feature Extraction from Input**
  - **Main goal is to quantify the degree of similarity between text pairs (input) i.e. convert text pairs into similarity scores (numeric values) so that Machine Learning algorithms can understand them**
  - **Note that Machine Learning algorithms can understand number values**
  - **We applied N-gram Overlap approach to compute similarity scores and transformed "data.csv" file into "features.csv"**

Uni-gram-Scores	Bi-gram-Score	Tri-gram-score	Four-gram-score	Five-gram-Score
0.6	0.5	0.33	0	0
0.6	0.25	0	0	0
0.4	0	0	0	0
0.5	0.33	0	0	0
0	0	0	0	0
1	1	1	1	0
1	1	1	1	1
0.33	0	0	0	0
0.67	0	0	0	0
0.17	0	0	0	0
0.12	0	0	0	0
0	0	0	0	0
0.75	0.57	0.33	0	0
0.57	0.31	0.08	0	0
0.25	0	0	0	0

- **Step 3: Represent "extracted features" and "output / label" into a format that Machine Learning algorithms can understand (normally saved as CSV file)**
  - **Feature.csv (WITH LABEL)**

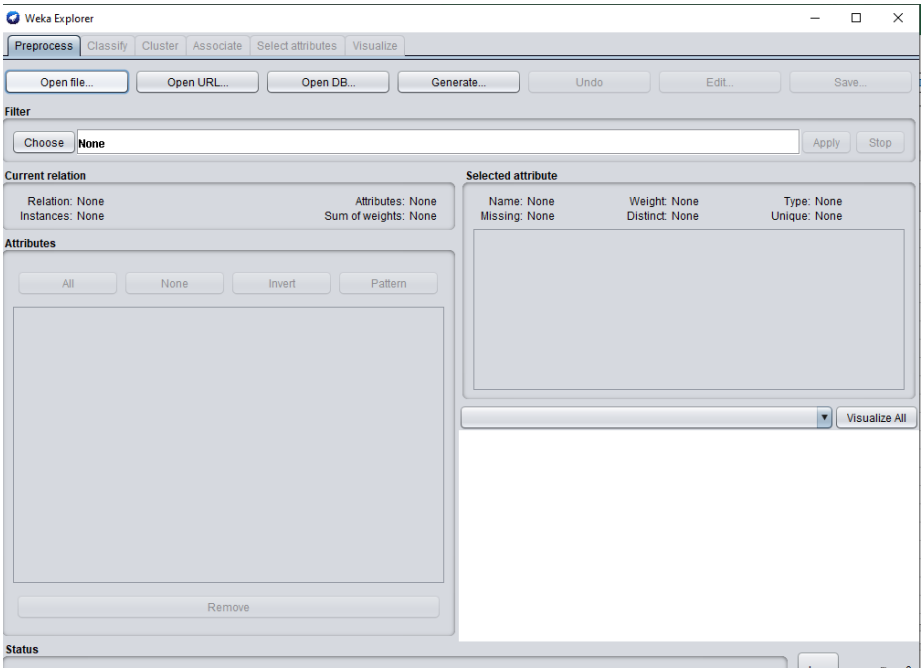
Uni-gram-Scores	Bi-gram-Score	Tri-gram-score	Four-gram-score	Five-gram-Score	Label
0.6	0.5	0.33	0	0	WD
0.6	0.25	0	0	0	PD
0.4	0	0	0	0	ND
0.5	0.33	0	0	0	ND
0	0	0	0	0	PD
1	1	1	1	0	WD
1	1	1	1	1	WD
0.33	0	0	0	0	PD
0.67	0	0	0	0	ND
0.17	0	0	0	0	ND
0.12	0	0	0	0	PD
0	0	0	0	0	PD
0.75	0.57	0.33	0	0	WD
0.57	0.31	0.08	0	0	WD
0.25	0	0	0	0	ND

- **Step 4: Use "features.csv" file to train and test Machine Learning algorithms**
  - **See Next Slides**

SLIDE

Example - Treating the Problem of Text Reuse Detection as Machine Learning Problem

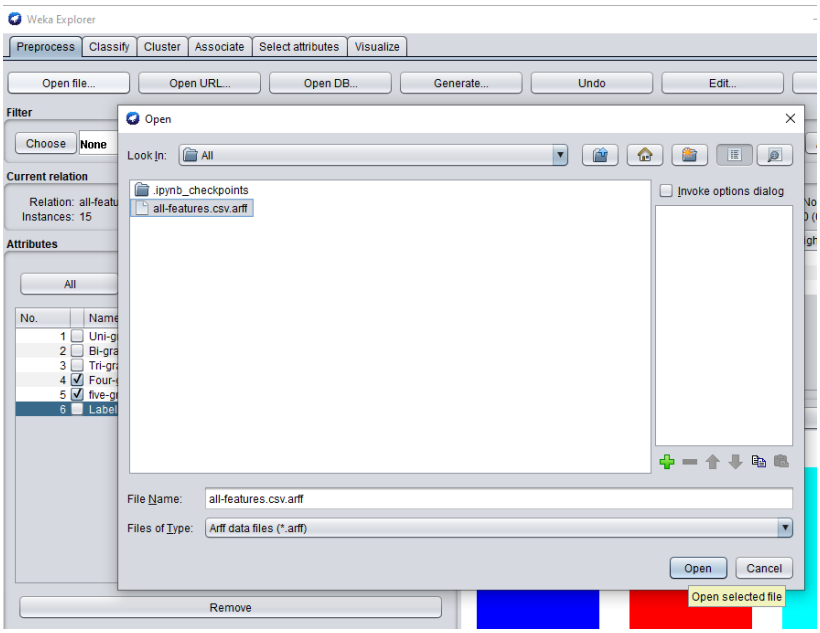
- **Ternary Classification using WEKA**
  - **Load features.csv**
    - **Click open**



SLIDE

Example - Treating the Problem of Text Reuse Detection as Machine Learning Problem (Cont…)

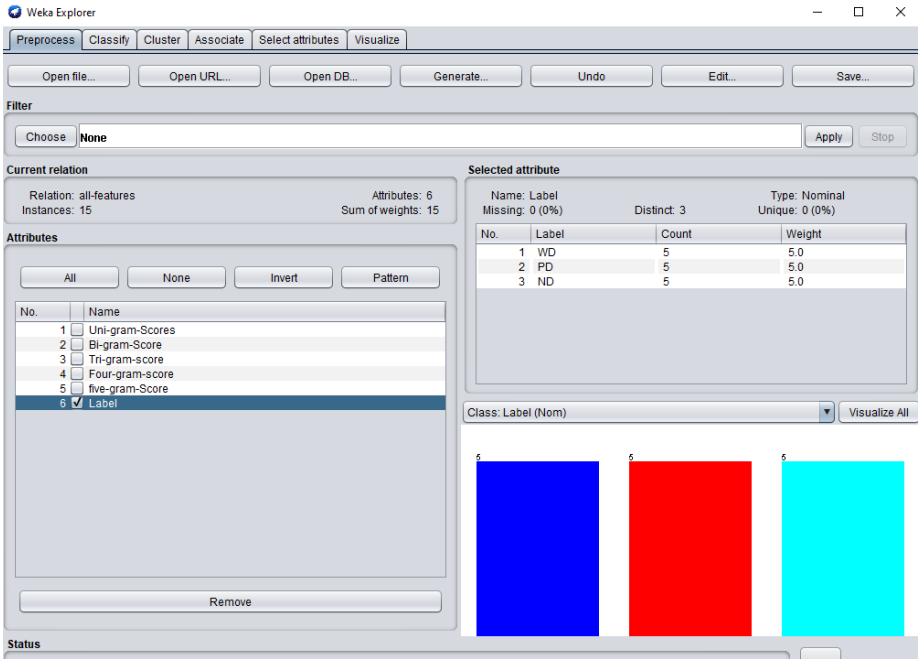
- **Ternary Classification using WEKA**
  - **Load features.csv**
    - **Select file by Giving Path**



SLIDE

Example - Treating the Problem of Text Reuse Detection as Machine Learning Problem (Cont..)

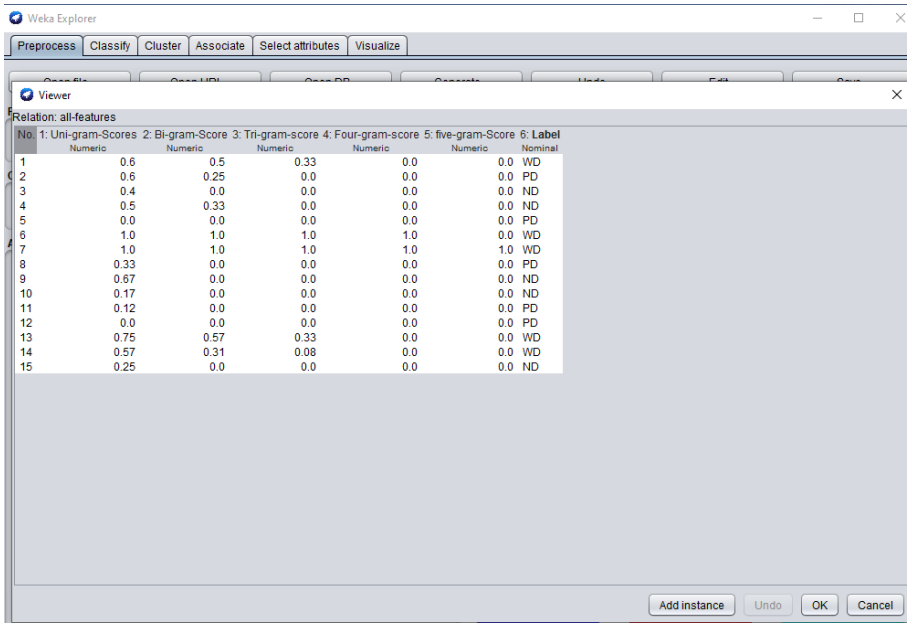
- **Ternary Classification using WEKA**
  - **Load features.csv**
    - **Click Label and see Number of Classes**



SLIDE

Example - Treating the Problem of Text Reuse Detection as Machine Learning Problem (Cont..)

- **Ternary Classification using WEKA**
  - **Load features.csv**
    - **Click Edit and see Data View in Weka**

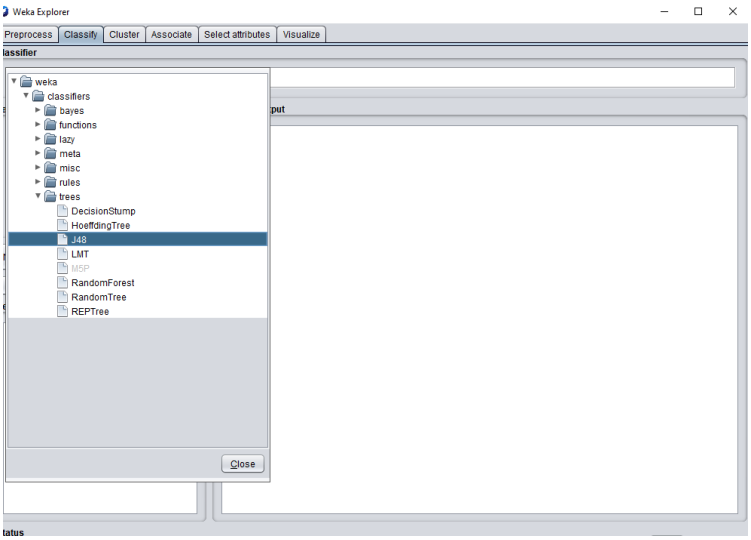


No.	1: Uni-gram-Scores	2: Bi-gram-Score	3: Tri-gram-score	4: Four-gram-score	5: five-gram-Score	6: Label
	Numeric	Numeric	Numeric	Numeric	Numeric	Nominal
1	0.6	0.5	0.33	0.0	0.0	WD
2	0.6	0.25	0.0	0.0	0.0	PD
3	0.4	0.0	0.0	0.0	0.0	ND
4	0.5	0.33	0.0	0.0	0.0	ND
5	0.0	0.0	0.0	0.0	0.0	PD
6	1.0	1.0	1.0	1.0	0.0	WD
7	1.0	1.0	1.0	1.0	1.0	WD
8	0.33	0.0	0.0	0.0	0.0	PD
9	0.67	0.0	0.0	0.0	0.0	ND
10	0.17	0.0	0.0	0.0	0.0	ND
11	0.12	0.0	0.0	0.0	0.0	PD
12	0.0	0.0	0.0	0.0	0.0	PD
13	0.75	0.57	0.33	0.0	0.0	WD
14	0.57	0.31	0.08	0.0	0.0	WD
15	0.25	0.0	0.0	0.0	0.0	ND

SLIDE

Example - Treating the Problem of Text Reuse Detection as Machine Learning Problem (Cont..)

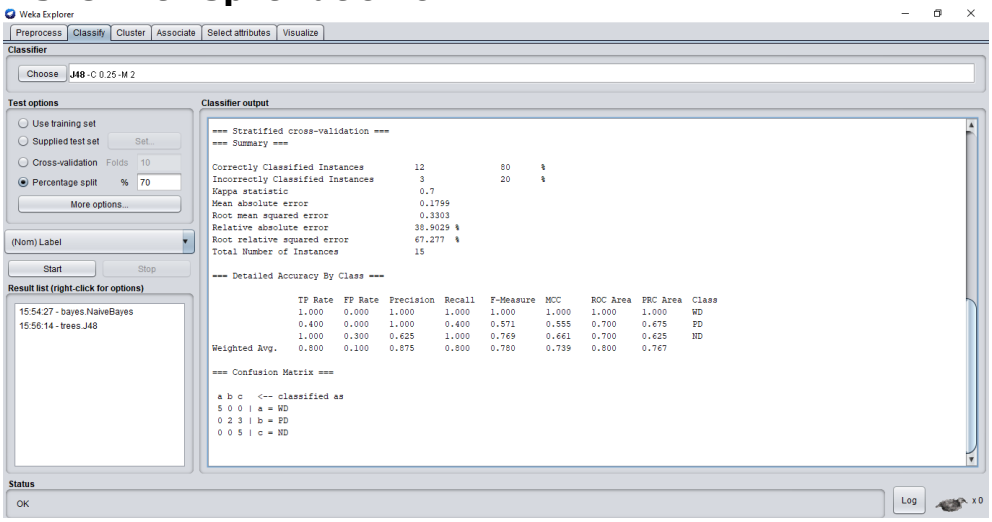
- **Ternary Classification using WEKA**
  - **Select J48**



SLIDE

Example - Treating the Problem of Text Reuse Detection as Machine Learning Problem (Cont..)

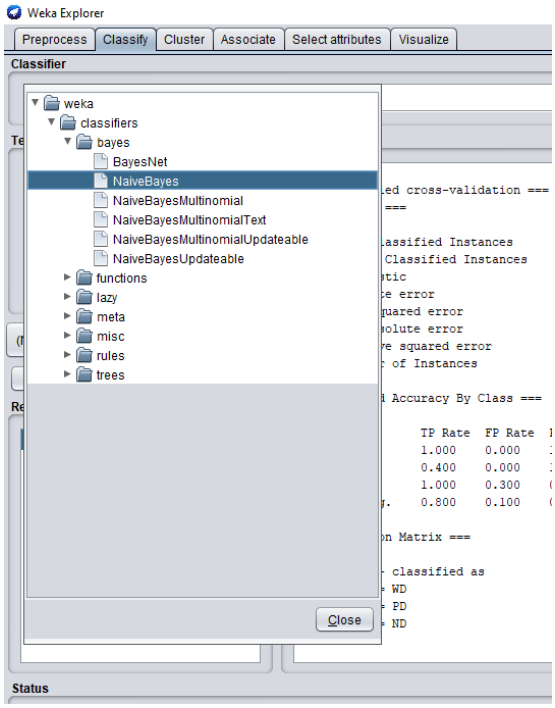
- **Ternary Classification using WEKA**
  - **Run J48 with Spilt ratio 70**



SLIDE

Example - Treating the Problem of Text Reuse Detection as Machine Learning Problem (Cont..)

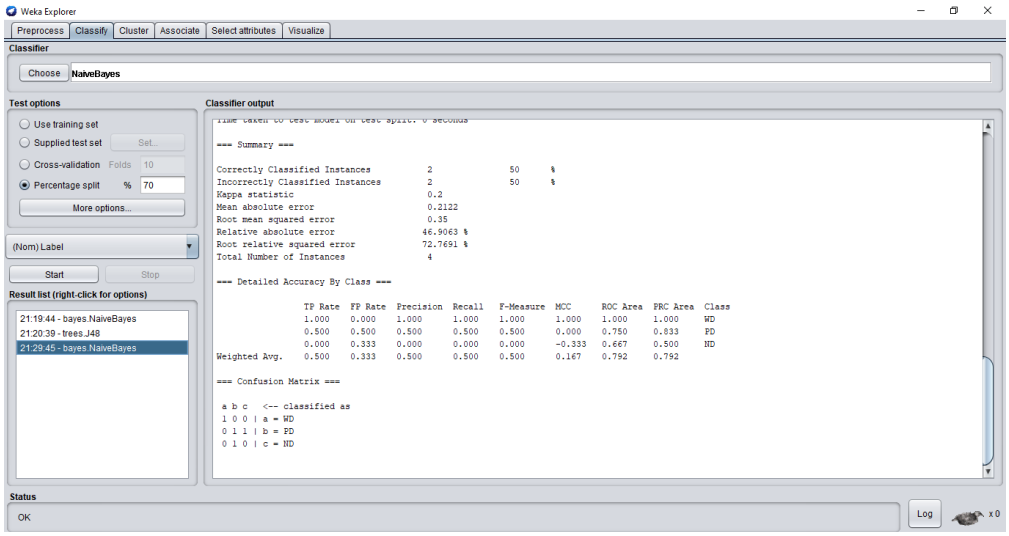
- **Ternary Classification using WEKA**
  - **Select Naive Bayes**



SLIDE

Example - Treating the Problem of Text Reuse Detection as Machine Learning Problem (Cont..)

- **Ternary Classification using WEKA**
  - **Run Naive Bayes with Spilt Ratio 70**



SLIDE

Results for Ternary Classification

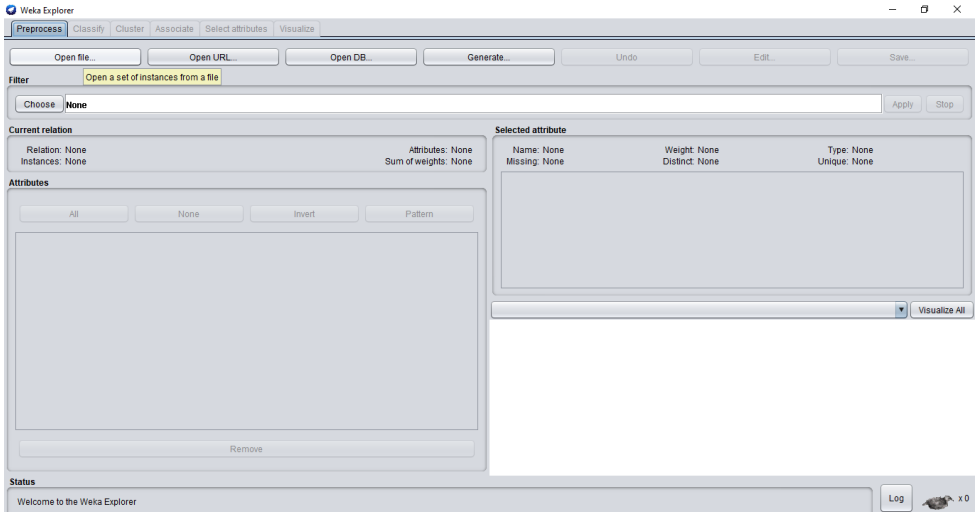
- **Results are reported for weighted average Precision, Recall and F_1 scores**

Machine Learning Algorithms	Results		
	Precision	**Recall**	**F_1-Measure**
Naïve Bayes	**0.500**	**0.500**	**0.500**
J48	**0.875**	**0.800**	**0.780**

SLIDE

Example - Treating the Problem of Text Reuse Detection as Machine Learning Problem

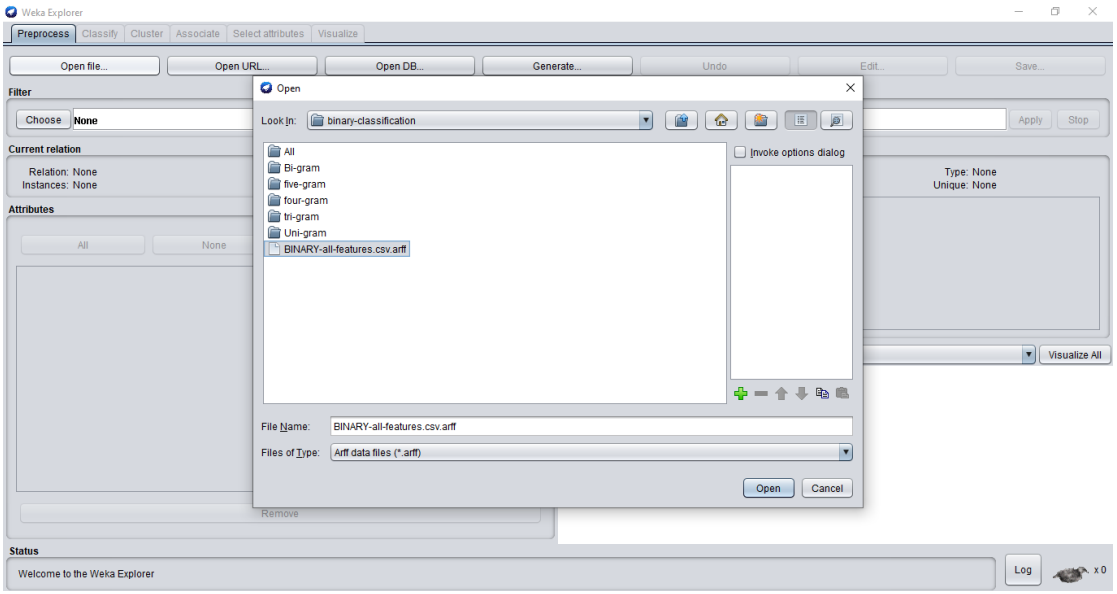
- **Binary Classification using WEKA**
  - **Load features.csv**
    - **Click Open**



SLIDE

Example - Treating the Problem of Text Reuse Detection as Machine Learning Problem

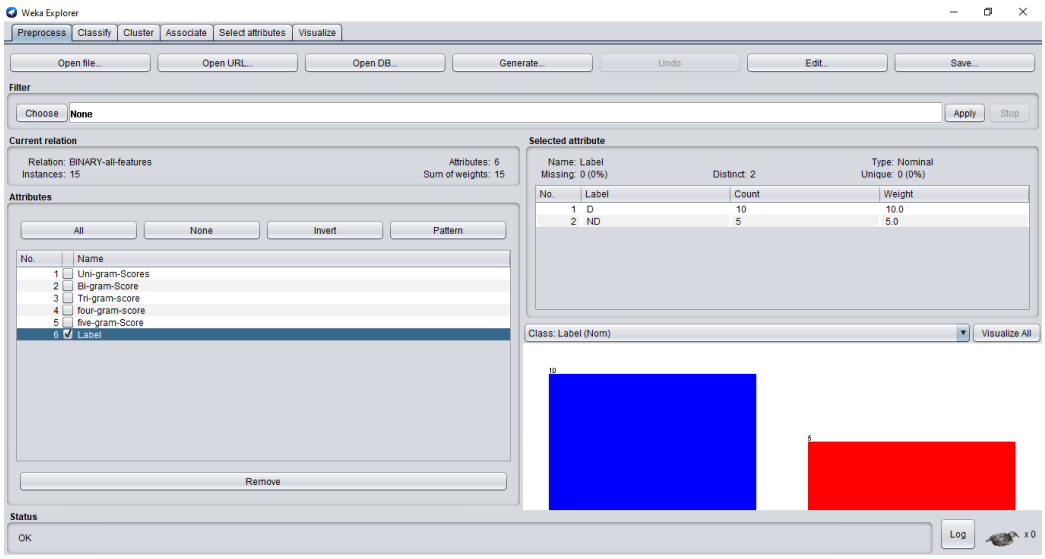
- **Binary Classification using WEKA**
  - **Load features.csv**
    - **Select file by Giving Path**



SLIDE

Example - Treating the Problem of Text Reuse Detection as Machine Learning Problem (Cont..)

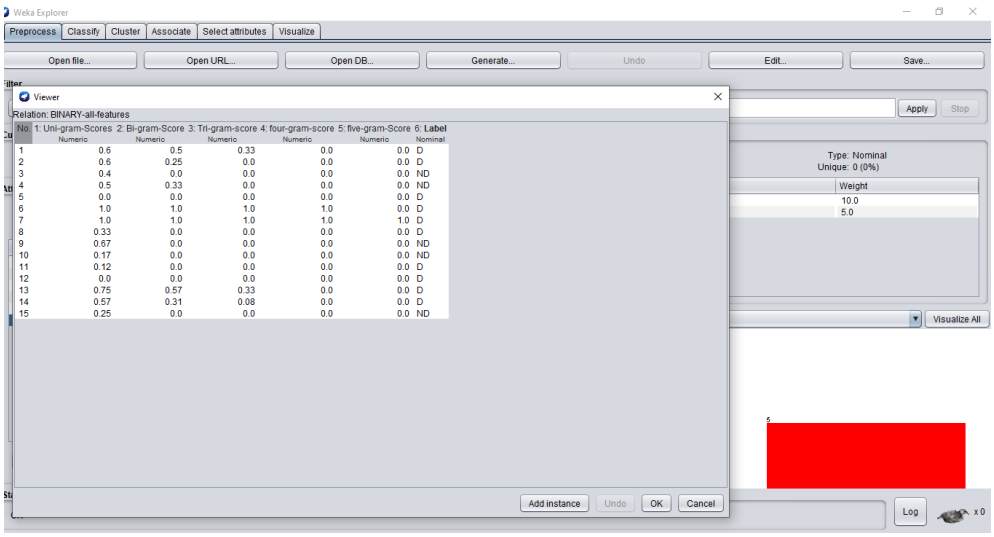
- **Binary Classification using WEKA**
  - **Load features.csv**
    - **Click Label and see Number of Classes**



SLIDE

Example - Treating the Problem of Text Reuse Detection as Machine Learning Problem (Cont..)

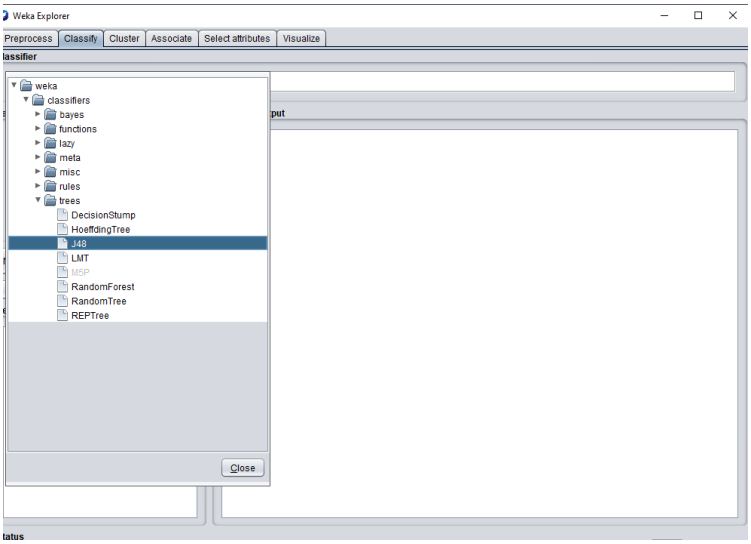
- **Binary Classification using WEKA**
  - **Load features.csv**
    - **Click Edit and see Data View in Weka**



SLIDE

Example - Treating the Problem of Text Reuse Detection as Machine Learning Problem (Cont..)

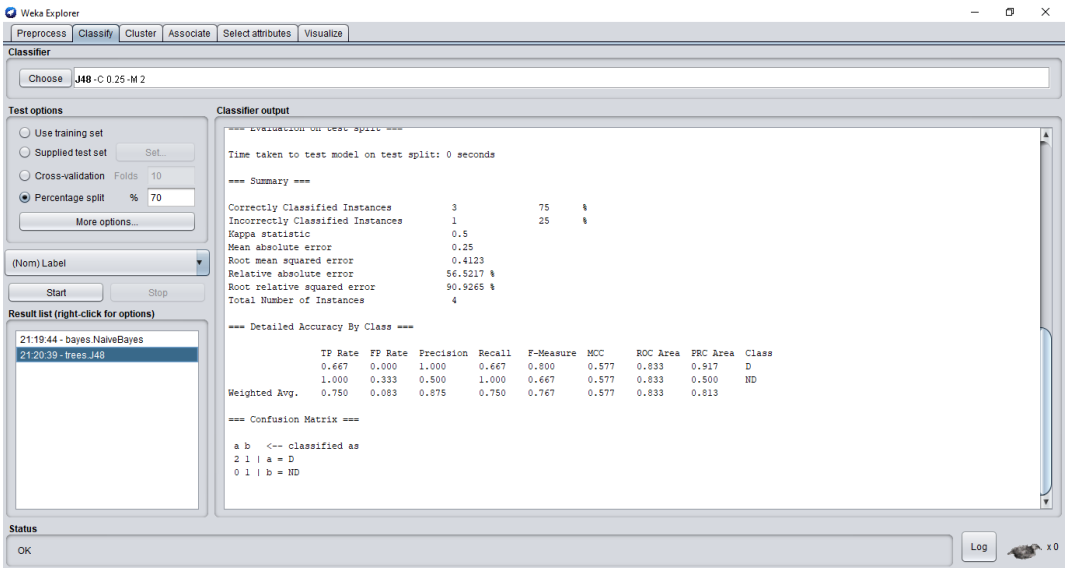
- **Binary Classification using WEKA**
  - **Select J48**



SLIDE

Example - Treating the Problem of Text Reuse Detection as Machine Learning Problem (Cont..)

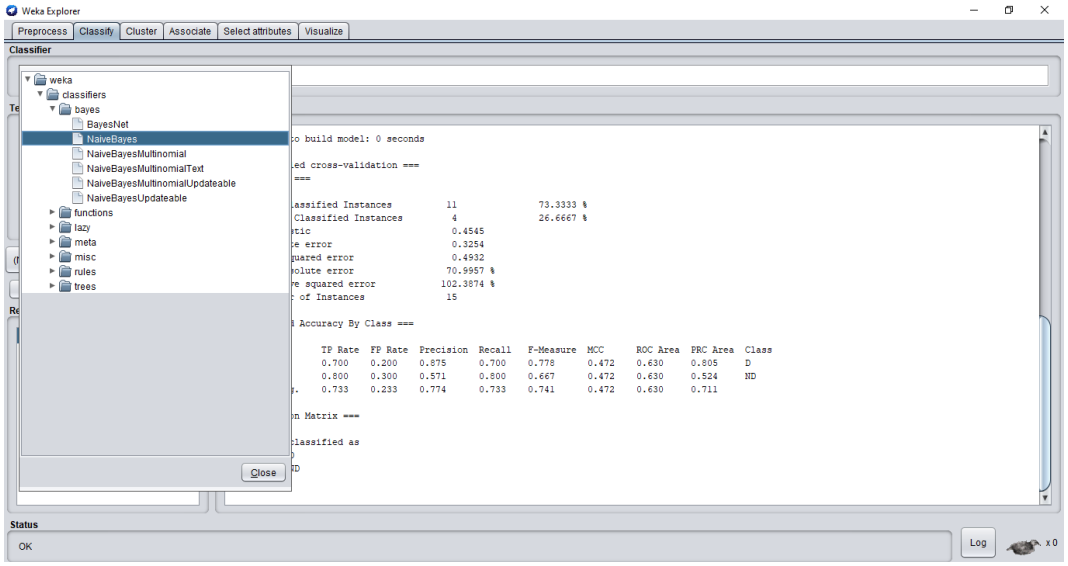
- **Binary Classification using WEKA**
  - **Run J48 with Spilt ratio 70**



SLIDE

Example - Treating the Problem of Text Reuse Detection as Machine Learning Problem (Cont..)

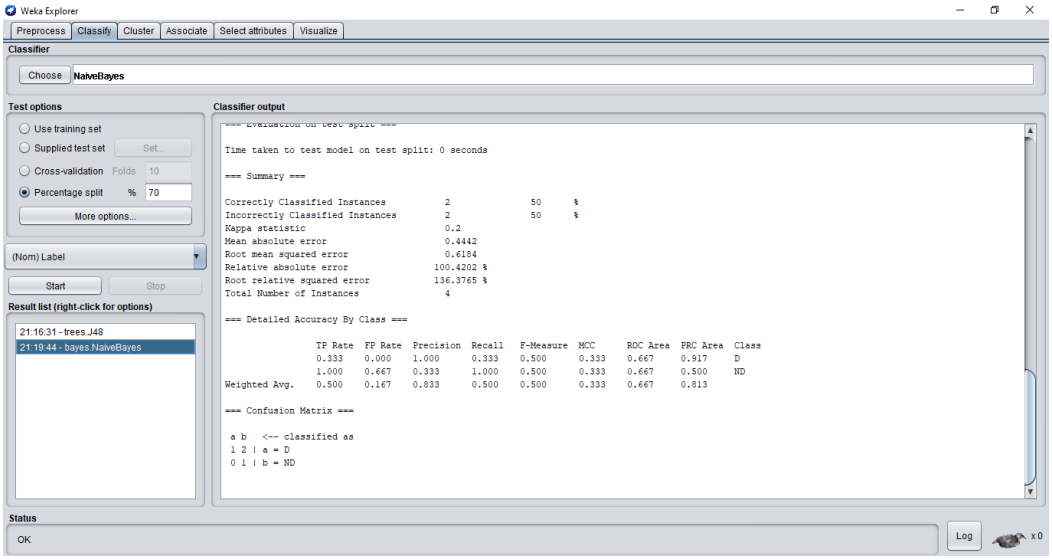
- **Binary Classification using WEKA**
  - **Select Naive Bayes**



SLIDE

Example - Treating the Problem of Text Reuse Detection as Machine Learning Problem (Cont..)

- **Binary Classification using WEKA**
  - **Run Naive Bayes with Spilt Ratio 70**



SLIDE

Results for Binary Classification

- **Results are reported for weighted average Precision, Recall and F_1 scores**

Machine Learning Algorithms	Results		
	Precision	Recall	F_1-Measure
Naïve Bayes	0.833	0.500	0.500
J48	0.875	0.750	0.767

SLIDE
Your Turn

- **Considering the toy dataset given in this lecture. Apply Longest Common Subsequence Approach to extract features (similarity scores between text pairs) from dataset. Convert the file into ARFF / CSV format. Run Naïve Bayes, J48 and two other Machine Learning algorithms from WEKA. Report Weighted Average Precision, Recall, and F_1 scores for all four Machine Learning algorithms in the form of table.**

SLIDE
Summary – Treating the Problem of Text Reuse / Plagiarism Detection as Machine Learning Problem – A Step by Step Example

- **To treat text reuse and plagiarism detection problem as a Supervised Text Classification task, we need to know following main things**
  - **Dataset**
    - **For supervised text classification task, dataset must be annotated**
  - **Technique(s)**
    - **To extract features from text pairs (input)**
      - **For text reuse and plagiarism detection the (feature extraction) techniques mostly aim to compute the similarity between text pairs i.e. features are similarity scores**
  - **Evaluation Measures**
    - **Mostly "weighted average" Precision, Recall and F_1 scores are used to evaluate the performance of text reuse and plagiarism detection systems**
  - **Machine Learning Toolkit(s)**
    - **A Machine Learning Toolkit is mainly a collection of Machine Learning algorithms**
    - **Two popular and widely used Machine Learning Toolkits are**
      - **WEKA (Java Programming Language)**
      - **Scikit-Learn (Python Programming Language)**

- **Machine Learning Algorithms**
  - **Machine Learning Algorithms that will be trained / texted on features (similarity scores) extracted from the dataset**
  - **For Supervised Text Classficiation task some of the Machine Learning Algorithms which have proven to be effective are**
    - **Naïve Bayes**
    - **Random Forest**
    - **Logistic Regression**
    - **Support Vector Machine**
    - **Multi-Layer Perceptron**
    - **AdaBoost**

- **Main Steps to Treat Text Reuse Detection as Supervised Text Classification Task**
  - **Step 1: Pre-process input (i.e. text pair)**
  - **Step 2: Feature Extraction from Input**
  - **Step 3: Represent "extracted features" and "output / label" into a format that Machine Learning algorithms can understand (normally saved as CSV file)**
  - **Step 4: Use "features" extracted in Stop 2 to train and test Machine Learning algorithms**

SLIDE

Summary – Introduction to Text Reuse and Plagiarism

- **Text Reuse**
  - **Text Reuse is the process of creating a new text (or document) using the existing one(s) and it is reported to be on rise in recent years due to easy access to large online digital repositories**
  - **Given a text pair (Text 1 and Text 2), Text 2 is said to be "Derived" from Text 1, if it is created using text from Text 1. On the other hand, Text 2 is said to be "Non-Derived" from Text 1, if it is interpedently written i.e. did not borrow text from Text 1**
  - **Text Reuse may occur at five levels: (1) Word level, (2) Phrasal level, (3) Sentence level, (4) Passage / Paragraph level and (5) Document level**
  - **Two main types of text reuse are: (1) Local Text Reuse - when amount of text reused is detected at sentence/passage level and (2) Global Text Reuse - when amount of text reused is detected at document level**
  - **Text Reuse can be: (1) Mono-lingual Text Reuse - when both the original and the reused text are in the same**

language and (2) Cross-lingual Text Reuse - when the original text is in one language and the reused text is in another language

- **Plagiarism**
  - **Plagiarism is defined as the unacknowledged reuse of text and in recent years it has been reported to be on rise. Consequently, plagiarism detection systems are routinely used by higher educational institutions to check students work for plagiarism**
  - **Given a text pair (source text and suspicious text), suspicious text is said to be "Plagiarized" from source text, if it is created using text from source text. On the other hand, suspicious text is said to be "Non-Plagiarized" from source text, if it is interpedently written i.e. did not borrow text from source text**
  - **There are three levels of Plagiarism: (1) Verbatim - the original text is reused as verbatim (word to word copy) or with minor modifications to create the plagiarized document, (2) Paraphrased Plagiarism - the original text is heavily altered (or paraphrased) to create the plagiarized document and (3) Plagiarism of Idea - the idea of the original text is reused without dependence on the words or form of the source**
  - **There are three main types of Plagiarism Cases: (1) Artificial Cases of Plagiarism - are generated by using Automatic Text Altering tools to obfuscate the source text for plagiarism, (2) Simulated / Manual Cases of Plagiarism - the original text is paraphrased by humans to create the cases of plagiarism and (3) Real Cases of Plagiarms - are those which occurred in the real world**
  - **Two main types of plagiarism detection are: (1) Intrinsic Plagiarism Detection - checking that the entire document (or all the passages) were written by one single author and (2) Extrinsic Plagiarism Detection - searching for the source(s) (or original text(s)) that were reused to create the suspicious document**
- **Data Annotation for Text Reuse Detection**
  - **Data Annotation**
    - **is the process of labeling data to make it usable for machine learning?**
    - **is performed by domain experts (humans – a.k.a. annotators / taggers / raters)**
    - **requires a lot of effort, time and cost**
  - **Main Steps to Create Benchmark Annotated Dataset**

- **Raw Data Collection**
  - **Data Source(s)**
  - **Cleaning of Data**
  - **Pre-processing of Data**
- **Annotation Process**
  - **Preparation of Annotation Guidelines**
  - **Annotations**
  - **Computing Inter-Annotator Agreement**
- **Corpus Standardization**

- **Methods for Text Reuse and Plagiarism Detection**
  - **Methods for Mono-lingual Text Reuse and Plagiarism Detection and be broadly categorized into: (1) Methods based on Content, (2) Methods based on Structure and (3) Methods based on Style**
  - **Methods for Cross-lingual Text Reuse and Plagiarism Detection and be broadly categorized into: (1) Methods based on Syntax, (2) Cross-Language Character N-Grams, (3) Methods based on Dictionaries, (4) Methods based on Parallel Corpora, (5) Methods based on Comparable Corpora, (6) Methods based on Word Embeddings and (7) Methods based on Deep Learning**
- **Evaluation Measures**
  - **Evaluation of Text Reuse / Plagiarism Detection Systems is carried out using Precision, Recall and F_1 measures**
  - **Note that in research papers (or thesis) mostly "weight average" Precision, Recall and F_1 scores are reported**
- **Treating the Problem of Text Reuse / Plagiarism Detection as Machine Learning Problem – A Step by Step Example**
  - **To treat text reuse and plagiarism detection problem as a Supervised Text Classification task, we need to know following main things**
    - **Dataset**
      - **For supervised text classification task, dataset must be annotated**
    - **Technique(s)**
      - **To extract features from text pairs (input)**
        - **For text reuse and plagiarism detection the (feature extraction) techniques mostly aim to compute the similarity between text pairs i.e. features are similarity scores**
    - **Evaluation Measures**
      - **Mostly "weighted average" Precision, Recall and F_1 scores are used to evaluate the performance of text reuse and plagiarism detection systems**

- **Machine Learning Toolkit(s)**
  - **A Machine Learning Toolkit is mainly a collection of Machine Learning algorithms**
  - **Two popular and widely used Machine Learning Toolkits are**
    - **WEKA (Java Programming Language)**
    - **Scikit-Learn (Python Programming Language)**
- **Machine Learning Algorithms**
  - **Machine Learning Algorithms that will be trained / texted on features (similarity scores) extracted from the dataset**
  - **For Supervised Text Classficiation task some of the Machine Learning Algorithms which have proven to be effective are**
    - **Naïve Bayes**
    - **Random Forest**
    - **Logistic Regression**
    - **Support Vector Machine**
    - **Multi-Layer Perceptron**
    - **AdaBoost**

- **Main Steps to Treat Text Reuse Detection as Supervised Text Classification Task**
  - **Step 1: Pre-process input (i.e. text pair)**
  - **Step 2: Feature Extraction from Input**
  - **Step 3: Represent "extracted features" and "output / label" into a format that Machine Learning algorithms can understand (normally saved as CSV file)**
  - **Step 4: Use "features" extracted in Stop 2 to train and test Machine Learning algorithms**