بِسْمِ اللهِ الرَّحْمٰنِ الرَّحِيْمِ

Research Methodology in I.T.

Lecture 01 – Introduction to Text Reuse and Plagiarism Detection

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

Instructor: Dr. Rao Muhammad Adeel Nawab

حضرت صوفی برکت علی صاحب رح

اے نوجوان!

نہ کہہ نہ لکھ، نہ کہہ نہ لکھ، نہ کہہ نہ لکھ

بہت کہا جا چکا بہت لکھا جا چکا، بہت کہا جا چکا بہت لکھا جا چکا، بہت کہا جا چکا بہت لکھا جا چکا

کر کے دکھا، کر کے دکھا، کر کے دکھا

دنیا تو تیرے کیے کو دیکھنا چاہتی ہے.

How to Work?


کام کرنا.


خوشی خوشی کام کرنا.


اللہ کو ساتھ لے کر خوشی خوشی کام کرنا.

آیت: إِيَّاكَ نَعْبُدُ وإِيَّاكَ نَسْتَعِينُ

ترجمہ: یا اللہ ہم تیری ہی عبادت کرتے ہیں.
اور تجھ ہی سے مدد مانگتے ہیں

Power of Effort & Dua

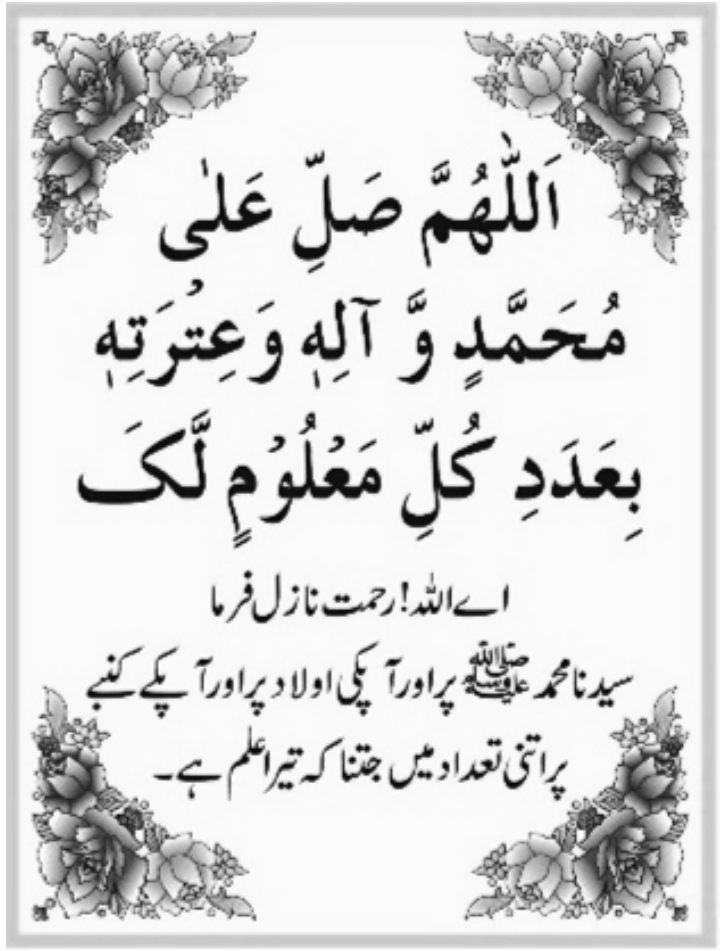
محنت کبھی نہیں ہارتی.

دعائیں ہوں تو کھوٹے سکے بھی چل جاتے ہیں.

دعا: اَهْدِنَا الصِّرَٰطَ الْمُسْتَقِيمَ صِرَٰطَ الَّذِينَ أَنْعَمْتَ عَلَيْهِم

ترجمہ: ہمیں سیدھی راہ دکھا ان لوگوں کی راہ جن پر تو نے انعام کیا.

Dua – Take Help from Allah Before Starting Any Task



اللَّهُمَّ خِرْ لِيْ وَاخْتَرْ لِي

سُبْحَانَكَ لَا عِلْمَ لَنَا إِلَّا مَا عَلَّمْتَنَا
إِنَّكَ أَنْتَ الْعَلِيمُ الْحَكِيمُ

رَبِّ اشْرَحْ لِي صَدْرِي
وَيَسِّرْ لِي أَمْرِي
وَاحْلُلْ عُقْدَةً مِنْ لِسَانِي
يَفْقَهُوا قَوْلِي

Dr. Rao Muhammad Adeel Nawab – About Me

PhD

University of Sheffield, UK

Assistant Professor

COMSATS University Islamabad, Lahore Campus

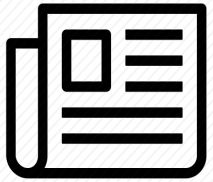
Group Lead

NLP Group, CUI, Lahore Campus

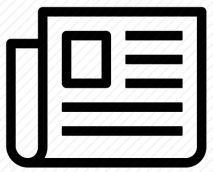
Publications

Research

36 International Publications
15 Impact Factor Journal Papers
21 International Conference/Workshop Papers



Citations & H-Index
258 Citations
H-Index = 10



Supervisions

10 PhD Students under supervision (in different countries)

40+ MPhil Students Supervised

About ME: Research Profile on Google Scholar

https://scholar.google.com/citations?user=fUn6DXYAAAAJ&hl=en

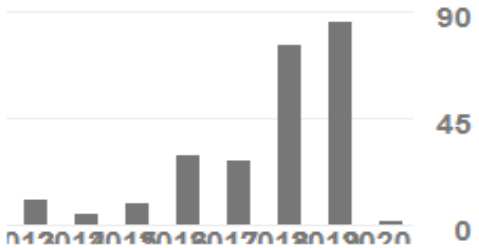
Dr. Rao Muhammad Adeel Nawab FOLLOW

Assistant Professor, Computer Science Department, COMSATS University Islamabad, Lahore Campus
Verified email at ciitlahore.edu.pk - Homepage

Natural Language ... Computational Lin... Data Science Artificial Intelligence

TITLE	CITED BY	YEAR
Multilingual author profiling on Facebook M Fatima, K Hasan, S Anwar, RMA Nawab Information Processing & Management 53 (4), 886-904	23	2017
Detecting Text Reuse with Modified and Weighted N-grams PC RMA Nawab, M Stevenson Proceedings of the First Joint Conference on Lexical and Computational ...	15 *	2012
Natural language descriptions of visual scenes: corpus generation and analysis MUG Khan, RMA Nawab, Y Gotoh Proceedings of the Joint Workshop on Exploiting Synergies between ...	14 *	2012
COUNTER: corpus of Urdu news text reuse M Sharjeel, RMA Nawab, P Rayson	13	2017

Cited by	All	Since 2015
Citations	258	231
h-index	10	9
i10-index	11	9



Co-authors EDIT

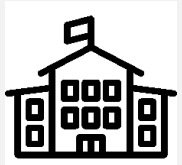
No co-authors

Dedication

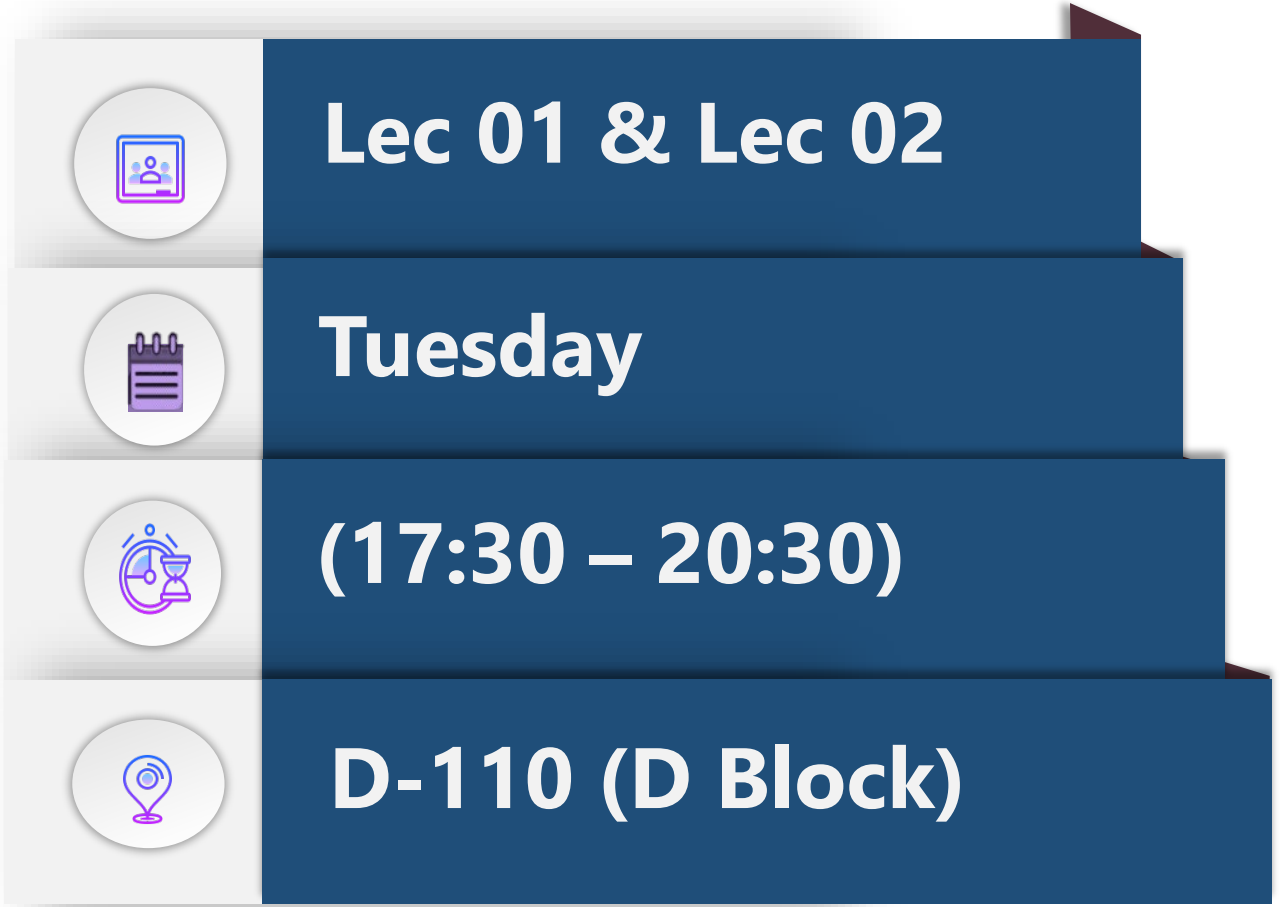
This course is dedicated to my most beloved and respectable PhD supervisor

Dr. Mark Stevenson
Department of Computer Science, University of Sheffield, UK

Course Details - Instructor

 **Instructor**	***Dr. Rao Muhammad Adeel Nawab***
 **Email**	***adeelnawab@cuilahore.edu.pk***
 **Office**	***Room no. 04, Faculty Block***
 **CUI Profile**	***https://lahore.comsats.edu.pk/Employees/74***

Course Details – Classes



Course Focus

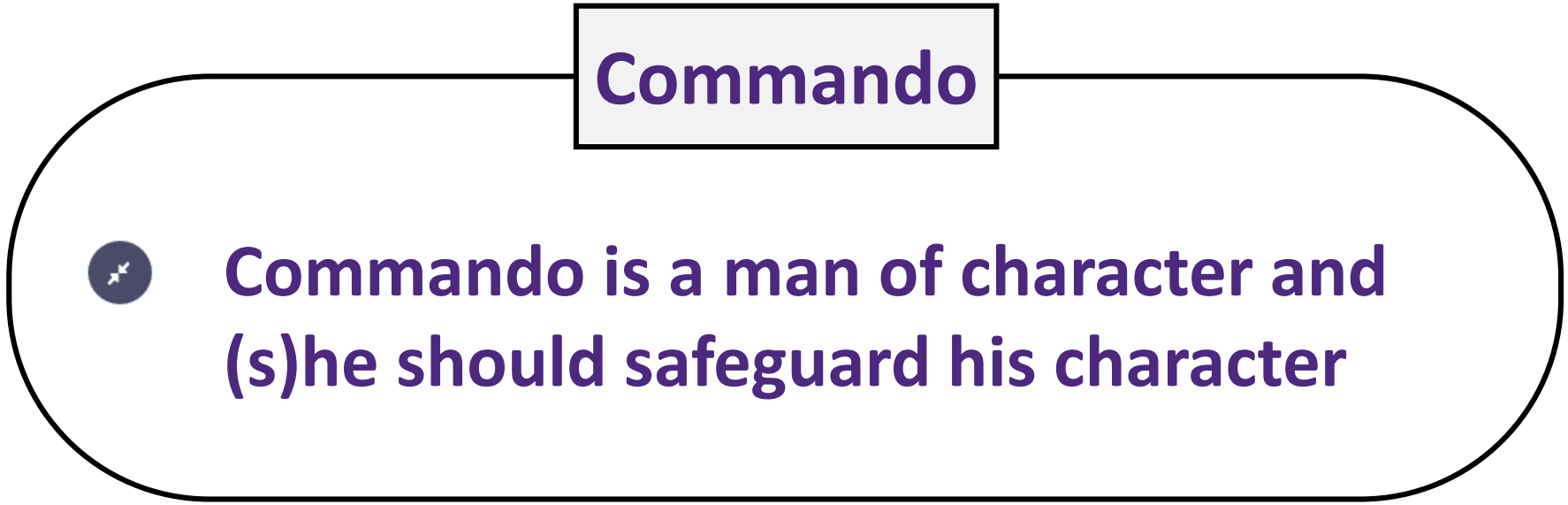
Mainly get Excellence in two things

- Become a great Human Being
- Become a great Researcher

Five Types of Training

- Police
- Elite
- Rangers
- Army
- Commando

Main Goal of a Course - Commando Training



Main Goal of a Course - Commando Training

Two Main Qualities of a Commando

1. محنت کبھی نہیں ہارتی.

100% Effort with Sincerity

2. دعائیں ہوں تو کھوٹے سکے بھی چل جاتے ہیں.

والدین اور اُستاد کی خدمت + ادب

Main Goal of a Course - Commando Training

Summary of Qualities in a Commando

عاجزی

ادب

- باادب بانصیب، بے ادب بے نصیب
- جو سیدھا بیٹھ نہیں سکتا، وہ سیدھا ہو نہیں سکتا
- جو ٹائم پر آ نہیں سکتا، وہ ٹائم پر کام نہیں کر سکتا

یہ راستہ اللہ کے فضل سے طے ہوتا ہے

اللہ کا فضل

جو ادب سے محروم ہے وہ
اللہ کے فضل سے محروم ہے

(جلال الدین رومی رضی اللہ عنہ)

Course Aims

1. To introduce essential concepts required to become a great human being and a great researcher
2. To develop skills to systematically teach any concept
3. To develop skills to carry out research using a template-based approach
4. To develop internet searching skills, both general and research specific

Course Aims

5 To develop skills to systematically search, read and summarize a research paper / thesis

6 To develop skills to systematically make a template-based outline of a research paper / thesis and then write it

7 To develop skills to systematically design an experiment

8 To develop skills to carry out research in such a way that students who carry out research become a Commando in life

Course Learning Outcomes (CLOs)

By the end of this course, the students should be able to

Understand what daily tasks are important to have a balanced personality

Understand how to systematically learn any concept

Understand how to search internet (both generic and research specific) to satisfy their information needs

Course Learning Outcomes (CLOs)

Read, write and design an experiment for a research paper / thesis using a template-based approach

Tell a coherent and connected story in a research paper / thesis

Carry out research in such a way that it enhances the self-learning abilities of students and create ability in them to cope up with the challenges of life

Useful Material

A Template-based Approach to Read a Research Paper
Download Link: https://ilmoirfan.com/a-template-based-approach-to-read-a-research-paper/

A Template-based Approach to Write a Research Paper
Download Link: https://ilmoirfan.com/a-template-based-approach-to-write-a-research-paper/

A Template-based Approach to Design an Experiment
Download Link: https://ilmoirfan.com/a-template-based-approach-to-design-an-experiment/

Useful Material Cont...

A Step by Step Approach to Learn Latex for Scientific Writing

Download Link: https://ilmoirfan.com/authors-page/latex/

A Template-based Approach to Write an Email

Download Link: https://ilmoirfan.com/authors-page/a-template-based-approach-to-write-an-email/

Little Efforts Daily Will Make You the Greatest

To systematically learn and get excellence in any concept / subject

- روز کا کام روز کریں

Importance of completing tasks on daily basis

- اک مہینے کا کھانا ایک دن میں نہیں کھایا جا سکتا، ایسے ہی ایک مہینے کا کام ایک دن میں نہیں ہو سکتا
- جو دن آپ کی زندگی سے چلا گیا اب واپس نہیں آئے گا
- آج کا کام آج ہی ہوسکتا ہے
- جو گزر گیا وہ آنا نہیں، آنے والے دن کا پتا نہیں، آج میدان جما ہے تو اپنا جوہر دکھاؤ

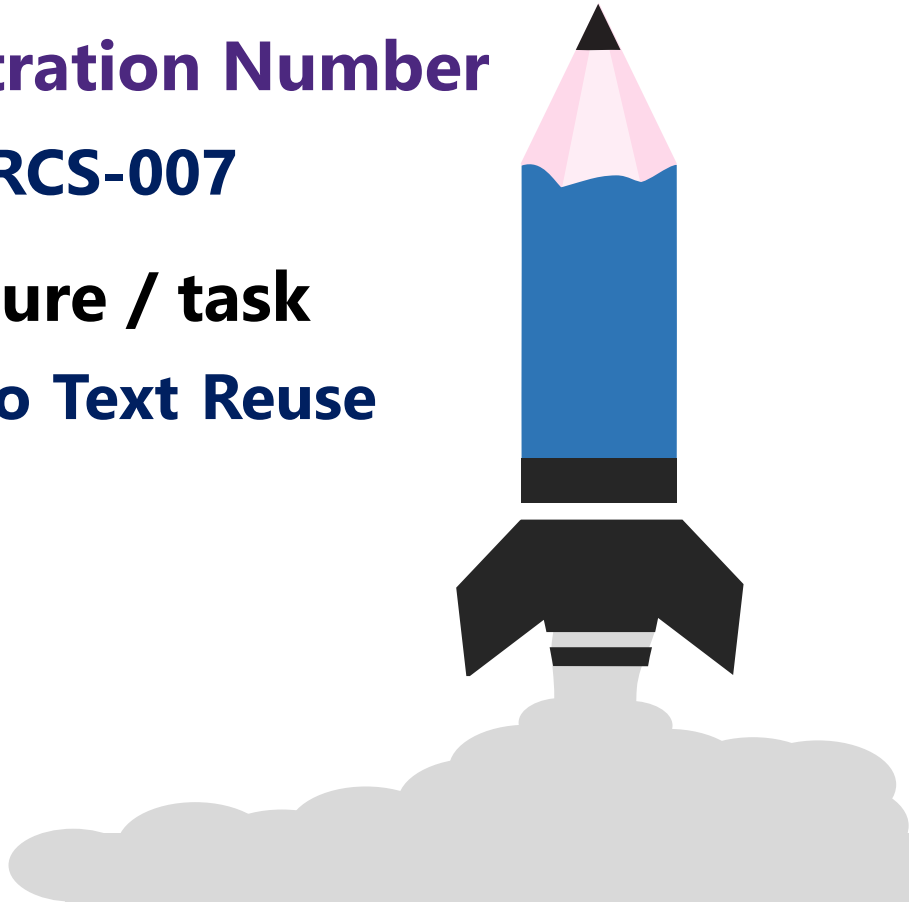
Instructions – To Do Tasks on Daily Basis

In order to facilitate the Course Instructor to monitor your work progress on daily basis, every student must follow the following steps:

Instructions – To Do Tasks on Daily Basis (Cont.)

1 **Create a folder on your Google Drive and share it's link with me on adeelnawab@cuilahore.edu.pk**

- **Name of Folder should be: Name – Registration Number**
  - » **For example: Muhammad Adeel – SP20-RCS-007**
- **Create a separate sub-folder for each lecture / task**
  - » **For example: Lecture 01 – Introduction to Text Reuse and Plagiarism**
- **Put your files in appropriate sub-folders**



Instructions – To Do Tasks on Daily Basis (Cont.)

2 **Update your files in Google Drive sub-folders on daily basis**

Very Important and Mandatory

The tasks mentioned in Your Turn slide(s) must be completed before the next lecture

Course Outline – Key Topics

- Introduction to Text Reuse and Plagiarism
- Learn How to Learn
- Learning is a Searching Problem
- Searching Offline and Online Sources of Knowledge and Skills
- A Template-based Approach to Analyze, Summarize and Document Search Results v1
- A Template-based Approach to Read a Research Paper v1
- A Template-based Approach to Design an Experiment
- A Template-based Approach to Write a Research Paper
- A Template-based Approach to Write a Research Thesis Proposal
- A Template-based Approach to Write a Research Thesis

Lecture Outline

1. Basics of Text Reuse
2. Basics of Plagiarism
3. Data Annotation for Text Reuse Detection
4. Methods for Text Reuse and Plagiarism Detection
5. Evaluation Measures
6. Treating the Problem of Text Reuse / Plagiarism Detection as Machine Learning Problem – A Step by Step Example

Basics of Text Reuse

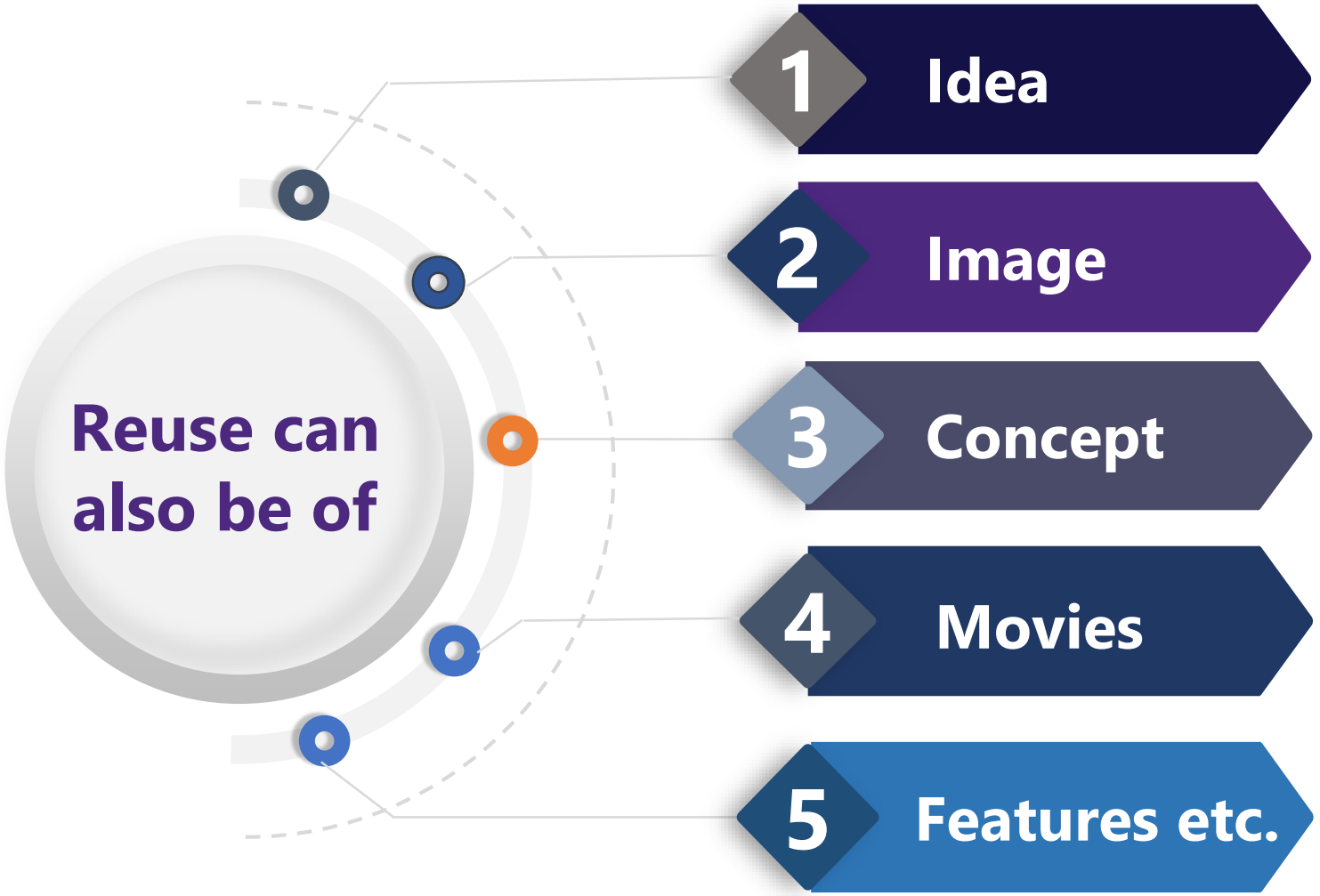
Text Reuse

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)**

Text Reuse


Reuse can also be of
1 Idea
2 Image
3 Concept
4 Movies
5 Features etc.

Text Reuse - Acceptable vs Non-Acceptable

1 Journalism

- Text reuse is a common practice
- Newspapers use text(s) provided by News Agencies to write newspaper articles

2 Plagiarism

- Unacknowledged text reuse is not acceptable

Text Reuse in Journalism

1 News Agency

» An organization that collects news items and distributes them to newspapers and broadcasters

2 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

Two Levels of Rewrite in Journalism

Derived vs Non-Derived

Derived

The Newspaper story was created borrowing the text(s) from News Agencies

Non-Derived

The Newspaper story is written independently and doesn't borrow any text from News Agencies

Three Levels of Re-write 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

Three Levels of Re-write in Journalism (Cont.)

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**

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
- Automatic text altering tools are making it easier to quickly modify text for reuse
- Freely available Machine Translation systems are helping people to easily even reuse text written in language that they don’t know

Text Reuse - Applications

1 Plagiarism Detection

Detecting unacknowledged reuse of text particularly in academia

2 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

3 Copyright infringement detection

Text Reuse Detection - Task

Given

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

Find

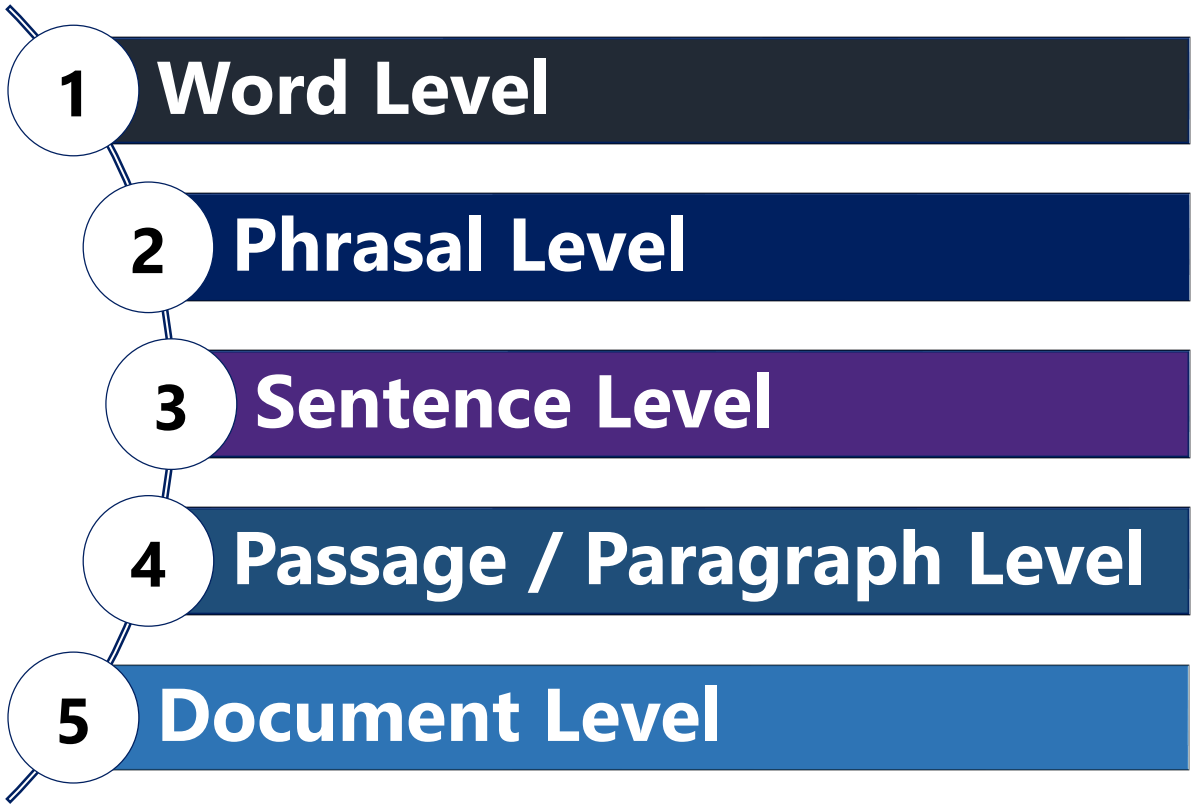
- How much text has been reused from Original (Text 1) to create Text 2 (output) i.e. goal is to identify the level of text reuse

Text Reuse – Input and Output

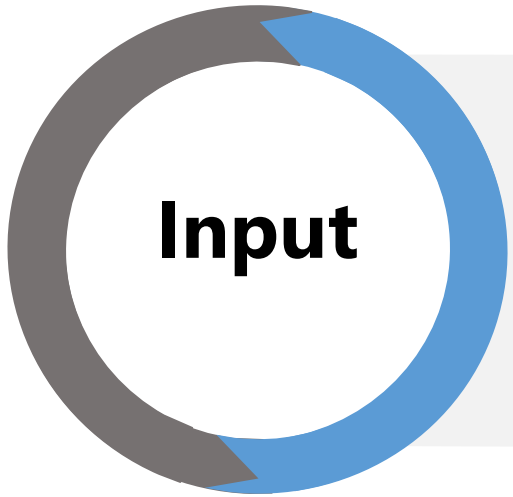
Input	**Text Pair (Text 1 and Text 2)**	
Output	**For two levels of text reuse** ▯ Derived ▯ Non-Derived	**For three levels of text reuse** ▯ Wholly Derived ▯ Partially Derived ▯ Non-Derived
Goal	**Identify the level of text reuse**	

Text Reuse - Granularity

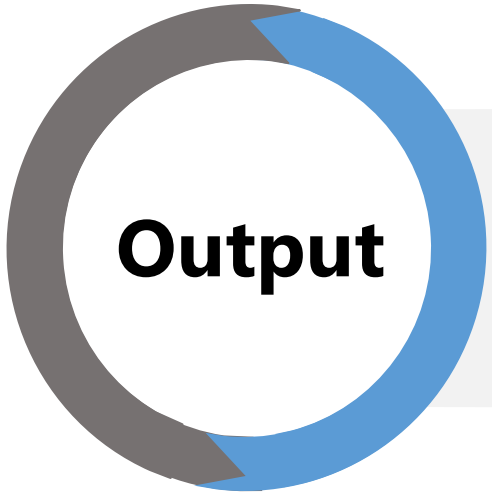
Text reuse may occur at five levels



Example 01 – Text Reuse at Word Level

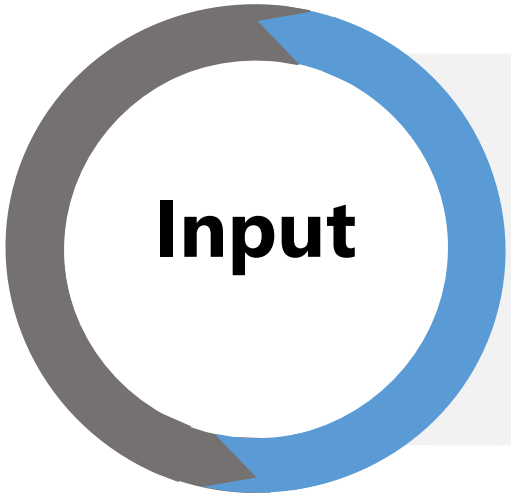


- Text 1
  - Meal
- Text 2
  - Food

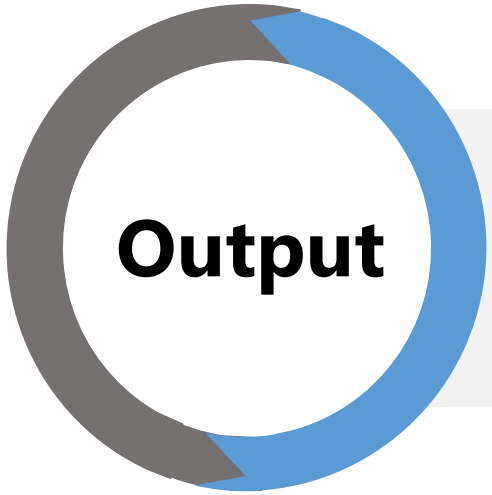


- Derived

Example 02 – Text Reuse at Word Level

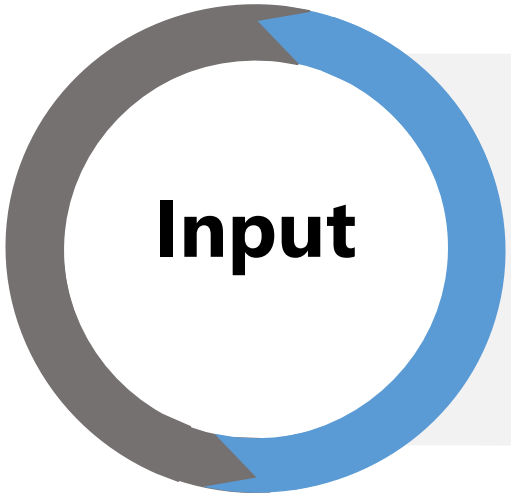


- Text 1
  - Meal
- Text 2
  - Butter

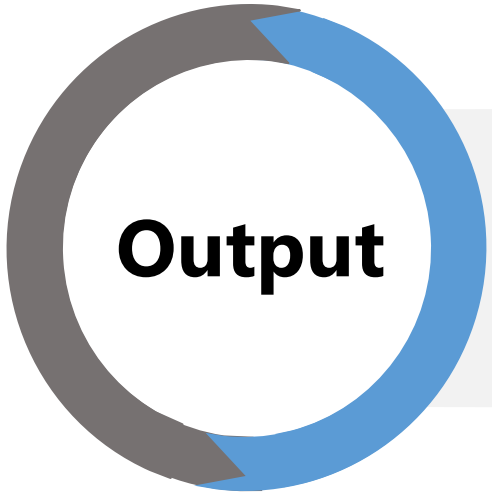


- Non-Derived

Example 03 – Text Reuse at Word Level

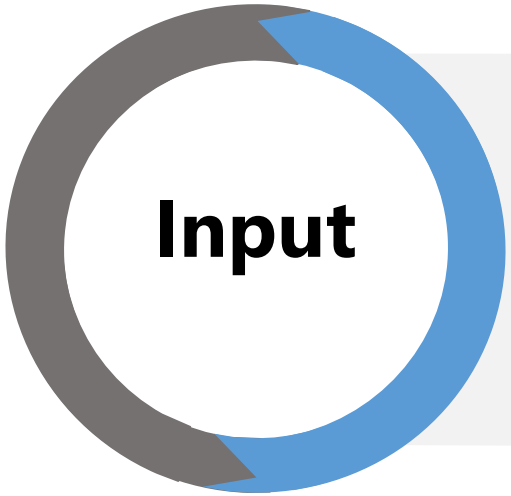


- Text 1
  - Like
- Text 2
  - Nice

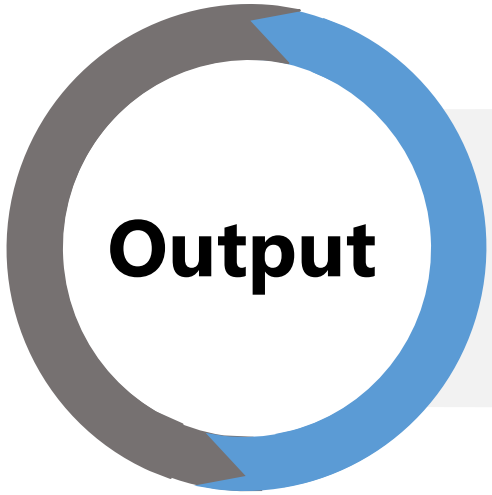


- Derived

Example 04 – Text Reuse at Word Level

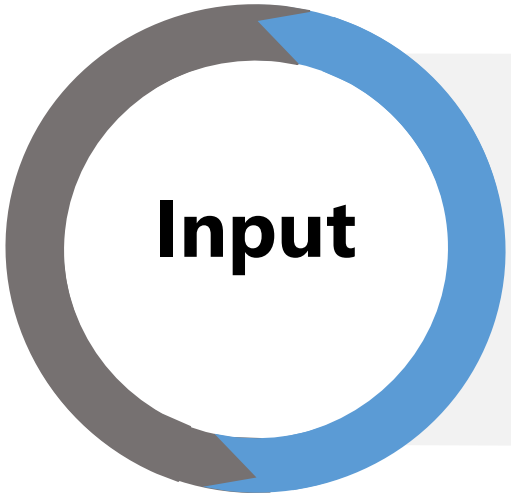


- Text 1
  - People
- Text 2
  - Audience

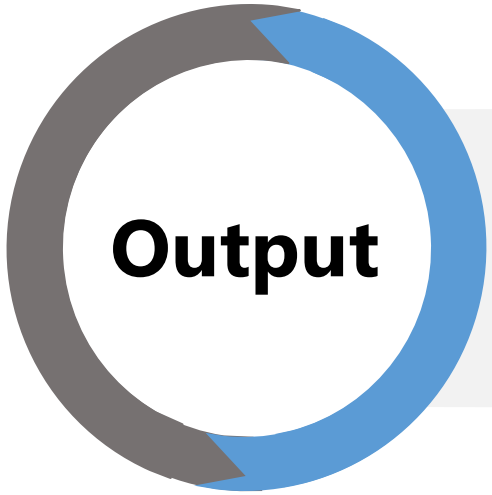


- Derived

Example 05 – Text Reuse at Word Level

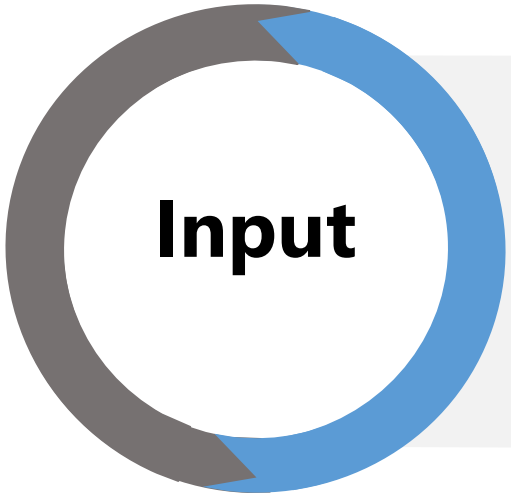


- Text 1
  - Dinner
- Text 2
  - Gathering

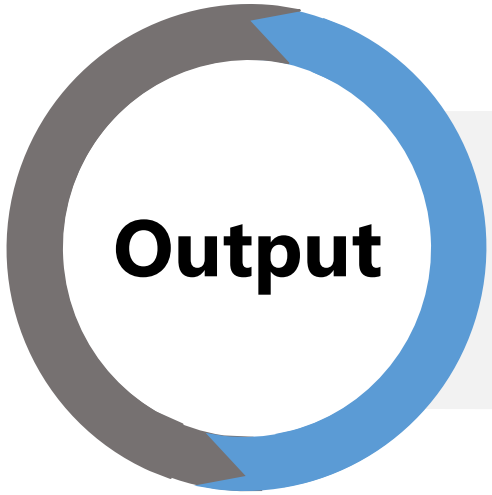


- Non-Derived

Example 06 – Text Reuse at Word Level

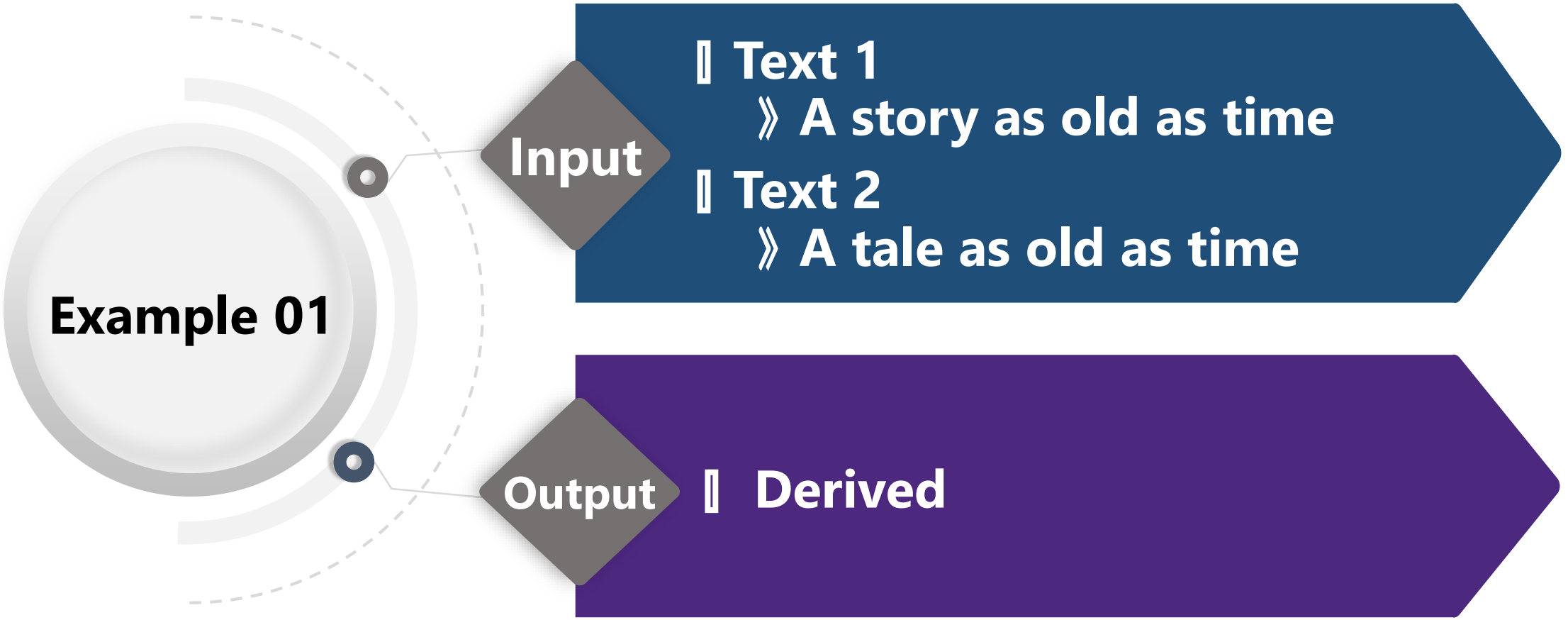


- Text 1
  - Dinner
- Text 2
  - Food

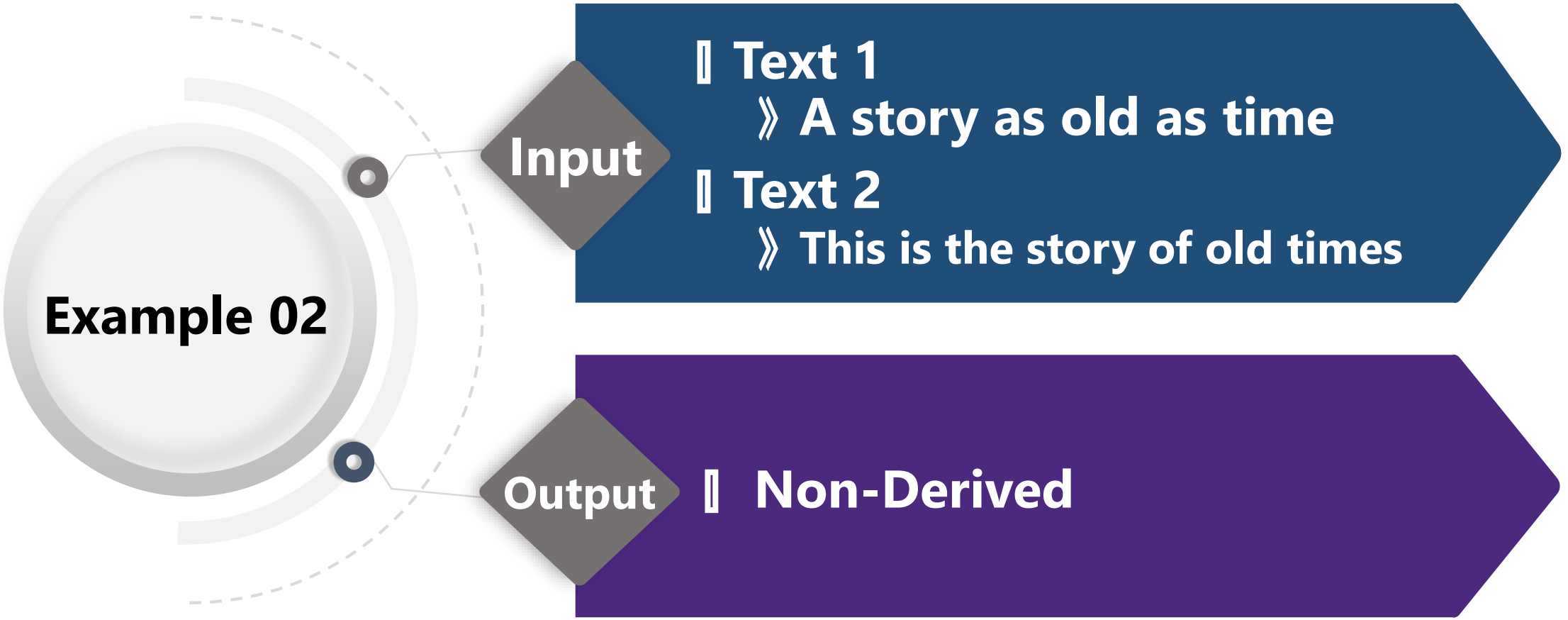


- Derived

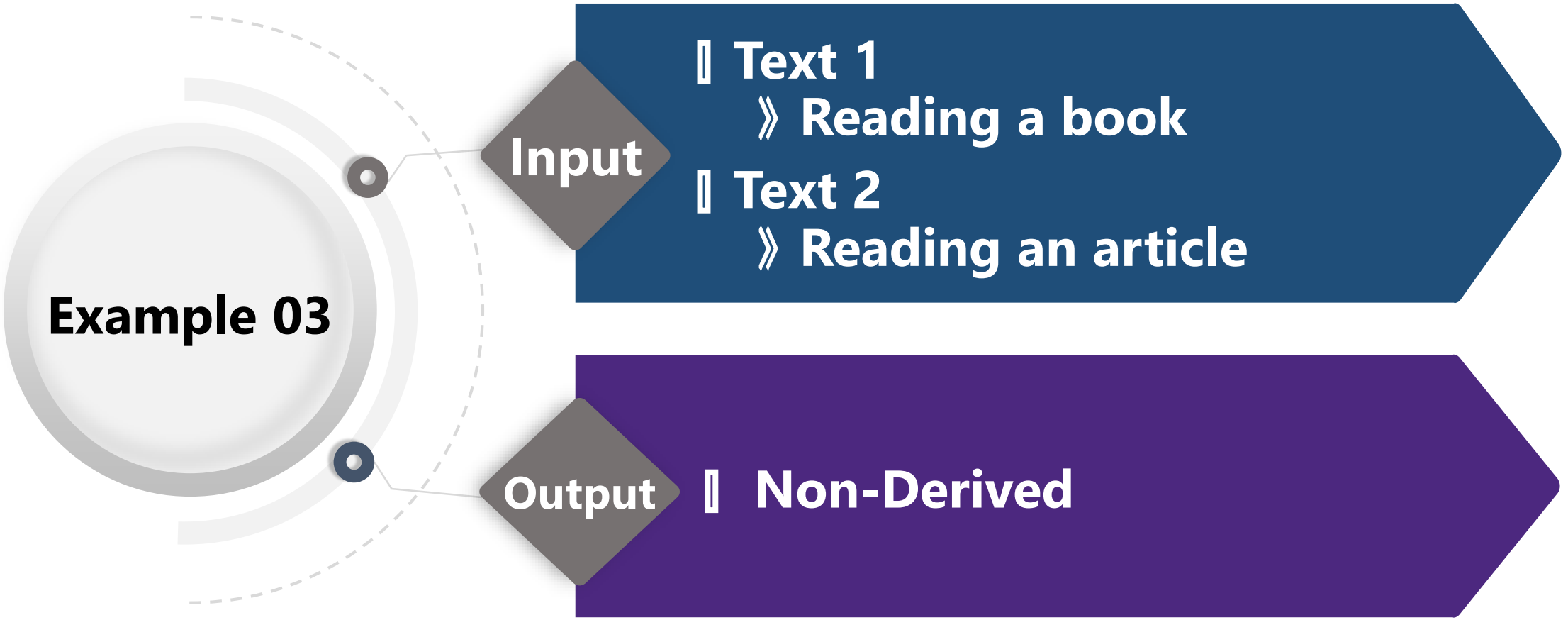
Example 01 – Text Reuse at Phrasal Level



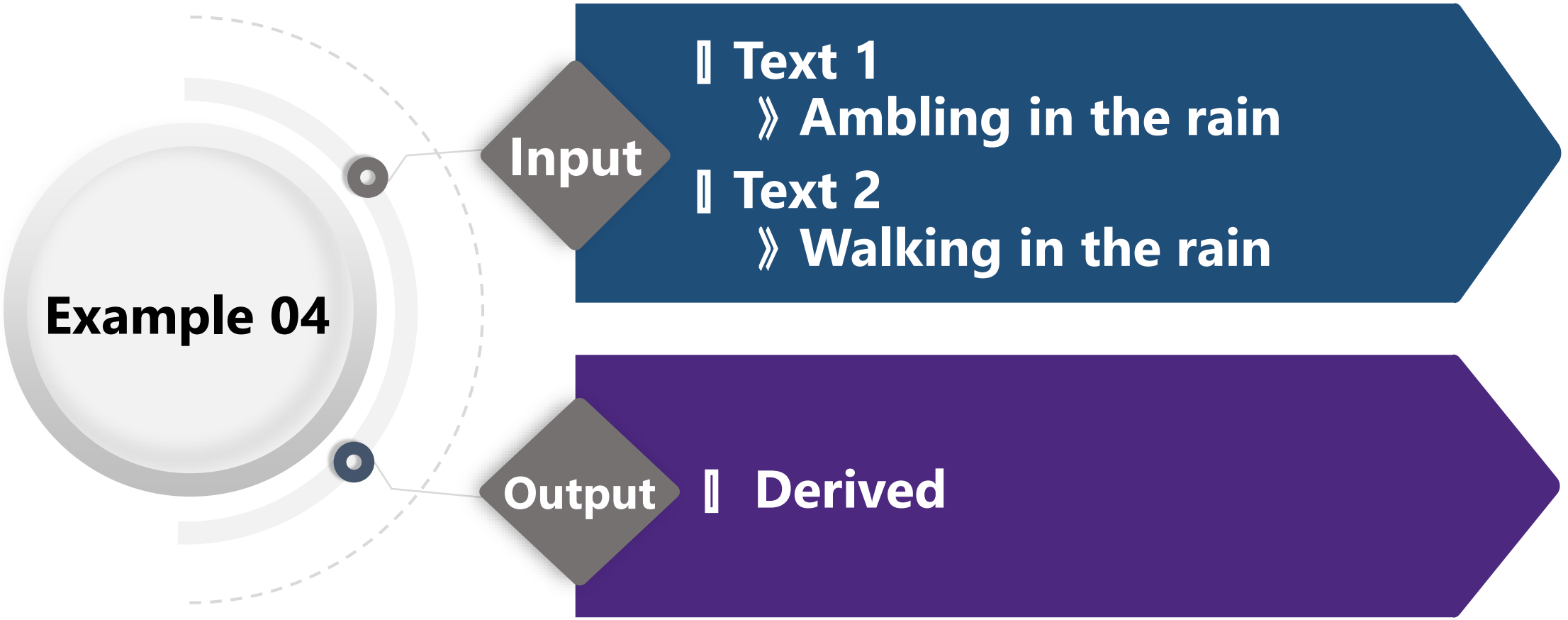
Example 02 – Text Reuse at Phrasal Level



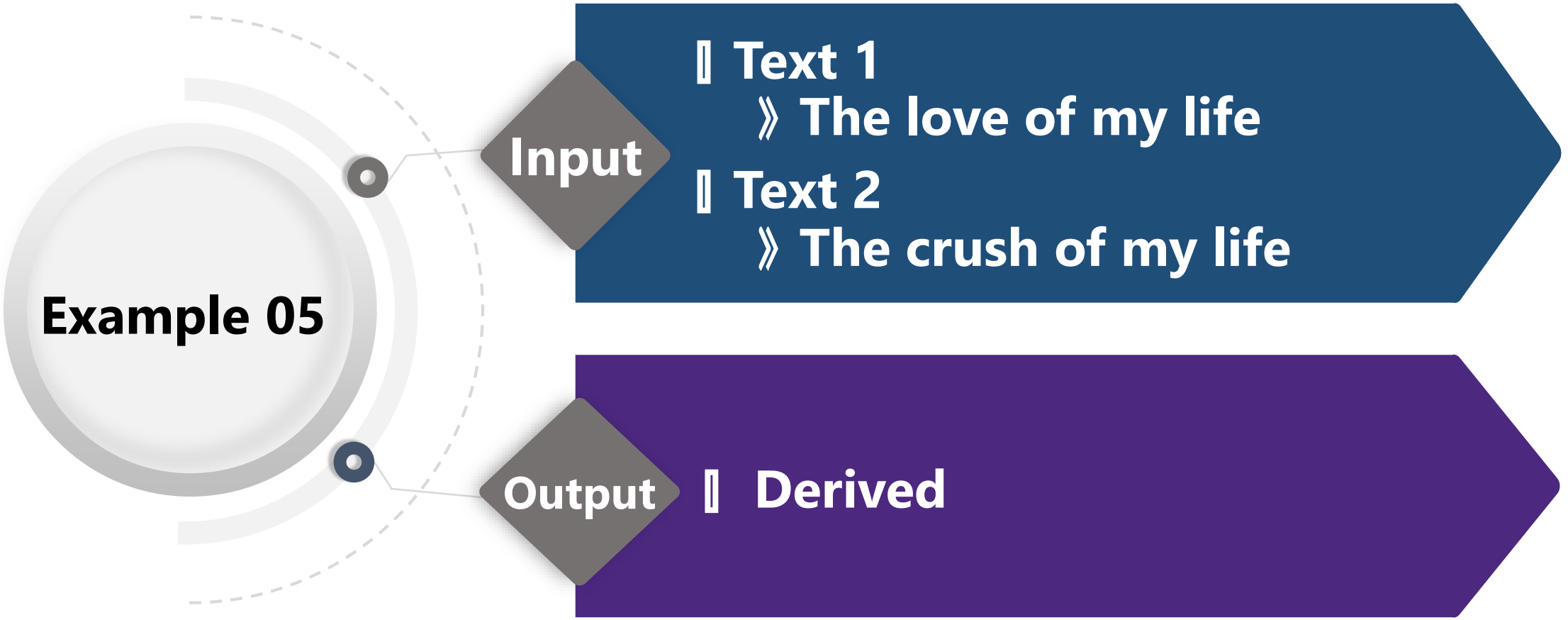
Example 03 – Text Reuse at Phrasal Level



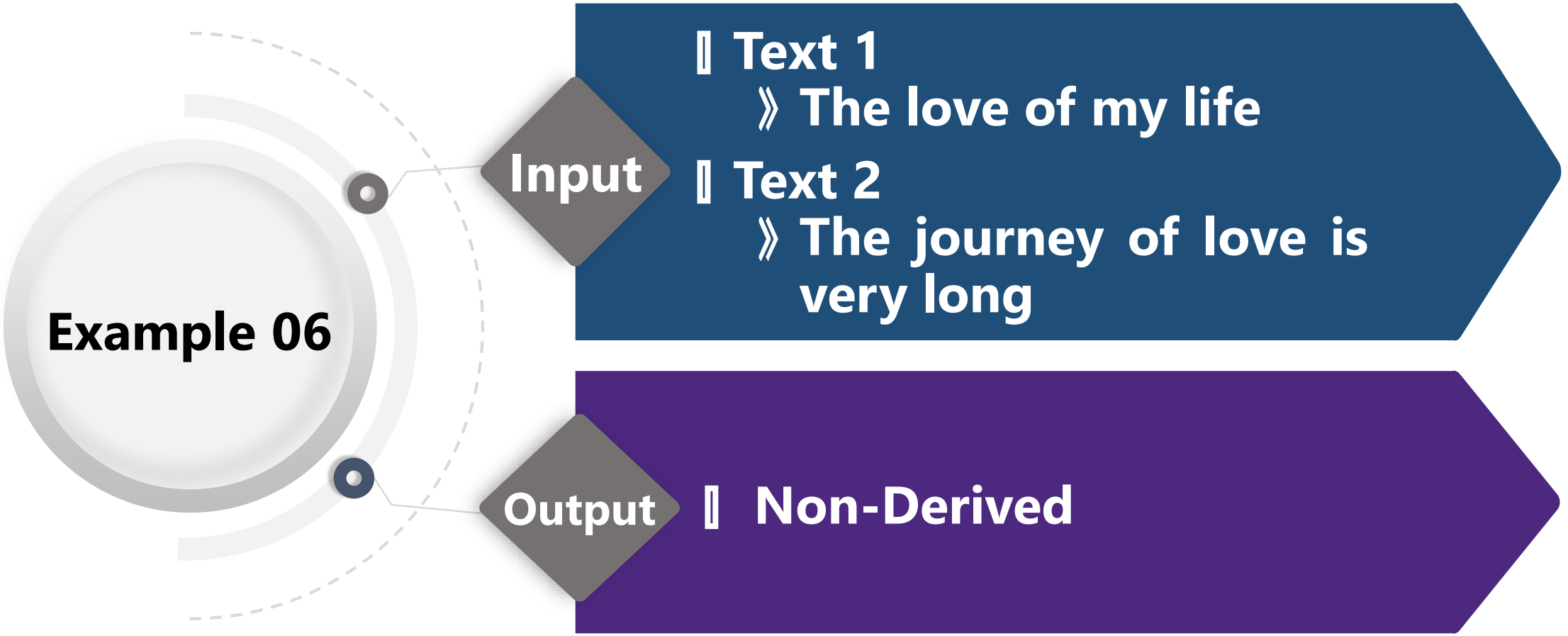
Example 04 – Text Reuse at Phrasal Level



Example 05 – Text Reuse at Phrasal Level



Example 06 – Text Reuse at Phrasal Level



Example 01 – Text Reuse at Sentence Level

Input

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

Output

- **Derived**

Example 02 – Text Reuse at Sentence Level

Input

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

Output

- **Non-Derived**

Example 03 – Text Reuse at Sentence Level

Input

- **Text 1**
  - » **Your car is nice**
- **Text 2**
  - » **I like your car**

Output

- **Derived**

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**

Example 05 – Text Reuse at Sentence Level

Input

- **Text 1**
  - **Balochistan successfully holds 3rd round of LG elections**
- **Text 2**
  - **Plots in the municipal elections, the PML-N won the third stage**

Output

- **Non-Derived**

Example 06 – 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**

Example 07 – 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**

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)

Output

Derived

Input (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

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

Output

Non-Derived

Input (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

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

Output

Derived

Input (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

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

Output

Non-Derived

Input (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

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

Output

Derived

Input (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

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

Output

Non-Derived

Input (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

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

Example 01 – Text Reuse at Document Level (Cont.)

Input (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.

Example 01 – Text Reuse at Document Level (Cont.)

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

Example 02 – Text Reuse at Document Level (Cont.)

Input (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, 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, hundreds of schools were destroyed by militants in Swat, once a tourist paradise of Swat was killed by terrorists

Example 02 – Text Reuse at Document Level (Cont.)

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

Example 03 – Text Reuse at Document Level

Input (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.

Example 03 – Text Reuse at Document Level (Cont.)

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

Example 04 – Text Reuse at Document Level

Input (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

Example 04 – Text Reuse at Document Level (Cont.)

Text Reuse - Types

Local Text Reuse vs Global Text Reuse

1 Local Text Reuse

When amount of text reused is detected at sentence / passage level

2 Global Text Reuse

When amount of text reused is detected at document level

Text Reuse - Types

Mono-lingual Text Reuse

- When both the original and the reused text are in the same language

VS

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 (e.g. Google Translator, Bign Translator etc.)
  - Manual Translation

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

Example - Cross-lingual Text Reuse

Input

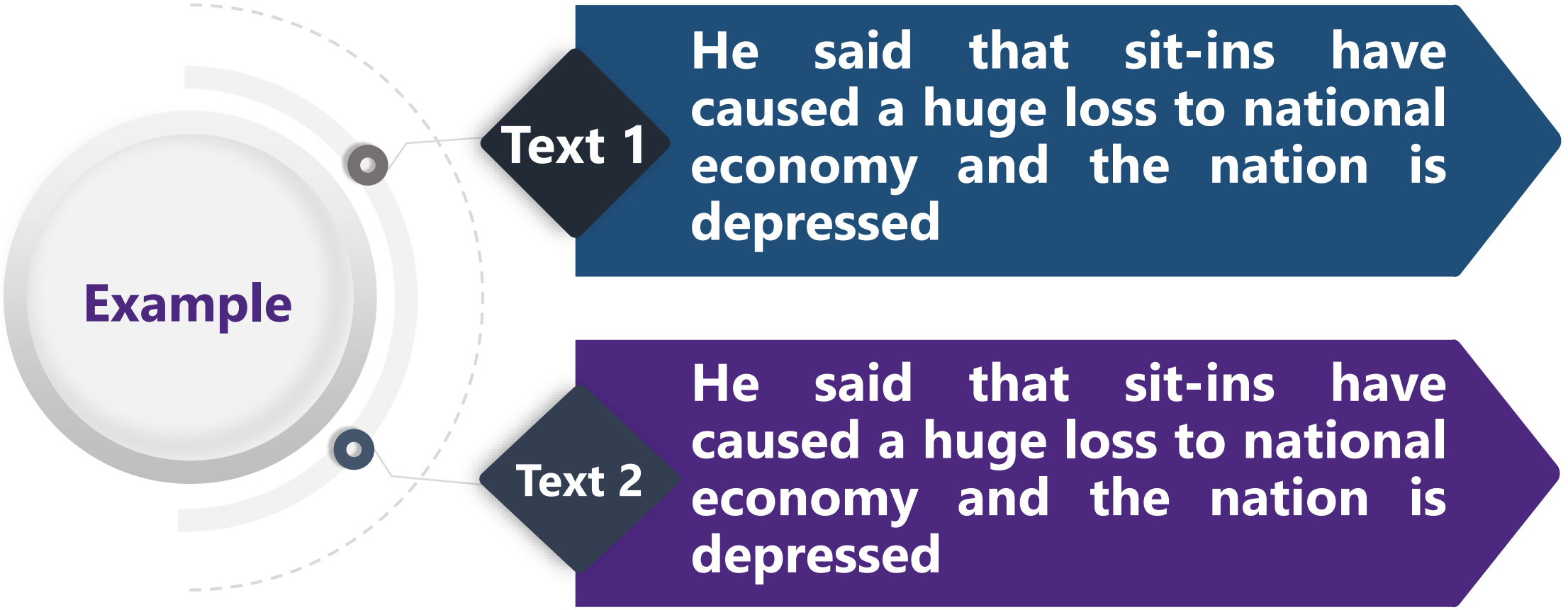
- **Text 1**
  - **When source and reused texts are in different languages**
    - **Source: A dog bites a man**
- **Text 2**
  - **ایک کتا ایک شخص کو کاٹتا ہے**

Output

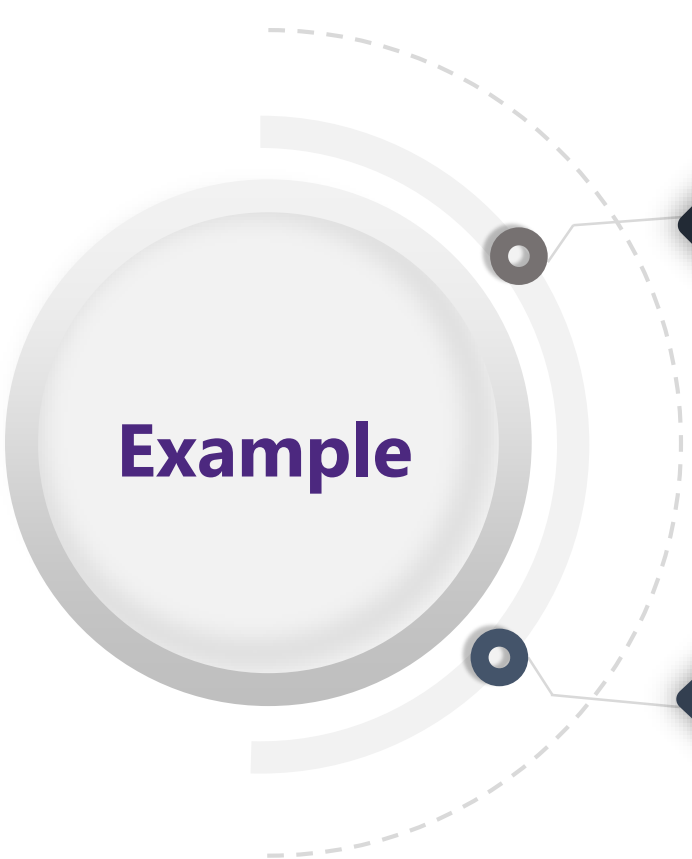
- **Derived**

Note **Both source and reused texts are in the different languages**

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



Example - Paraphrased Copy (Mono-lingual Settings)



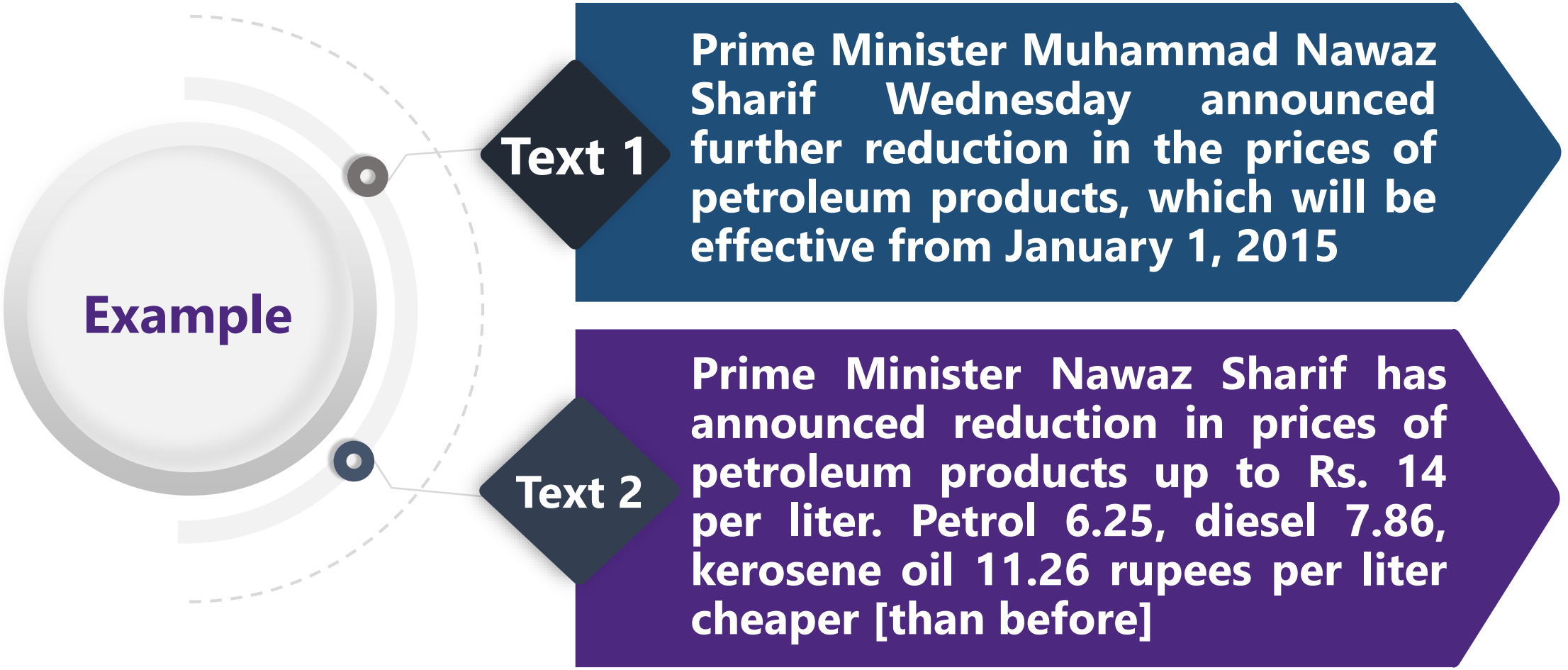
Text 1

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

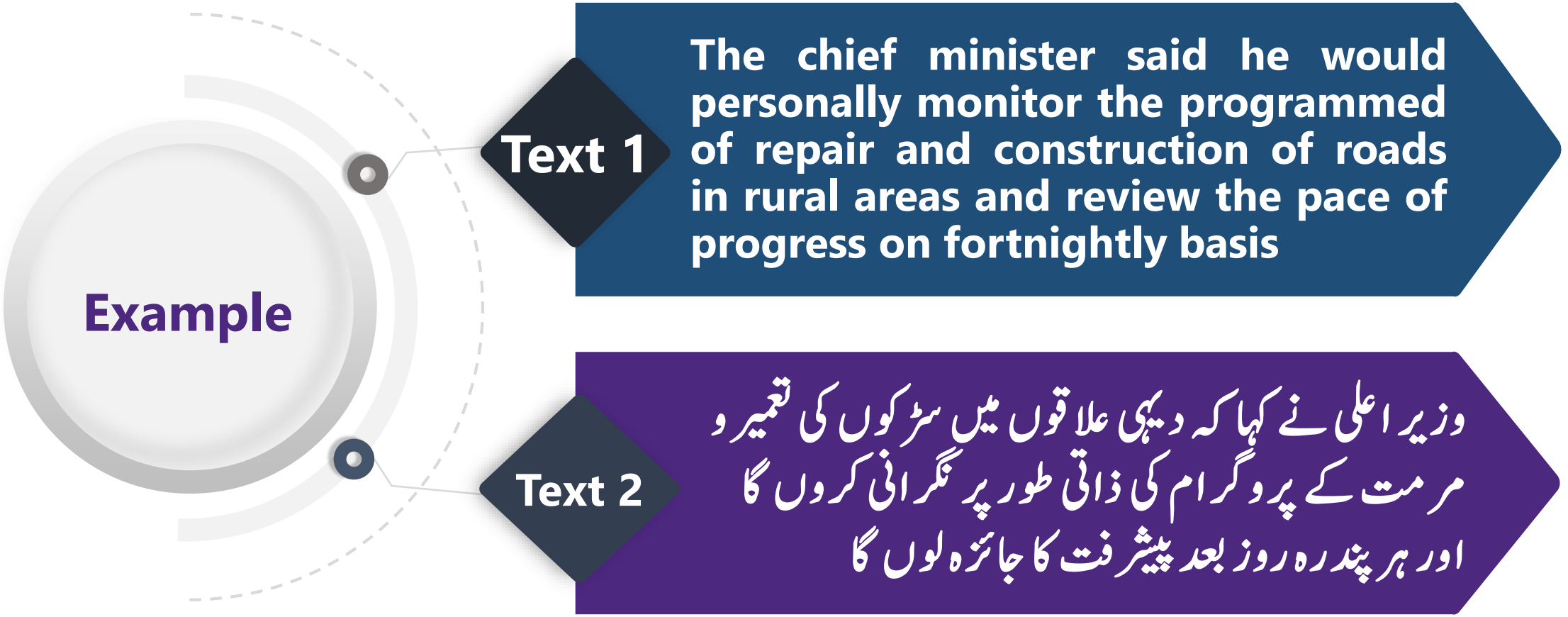
Text 2

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.

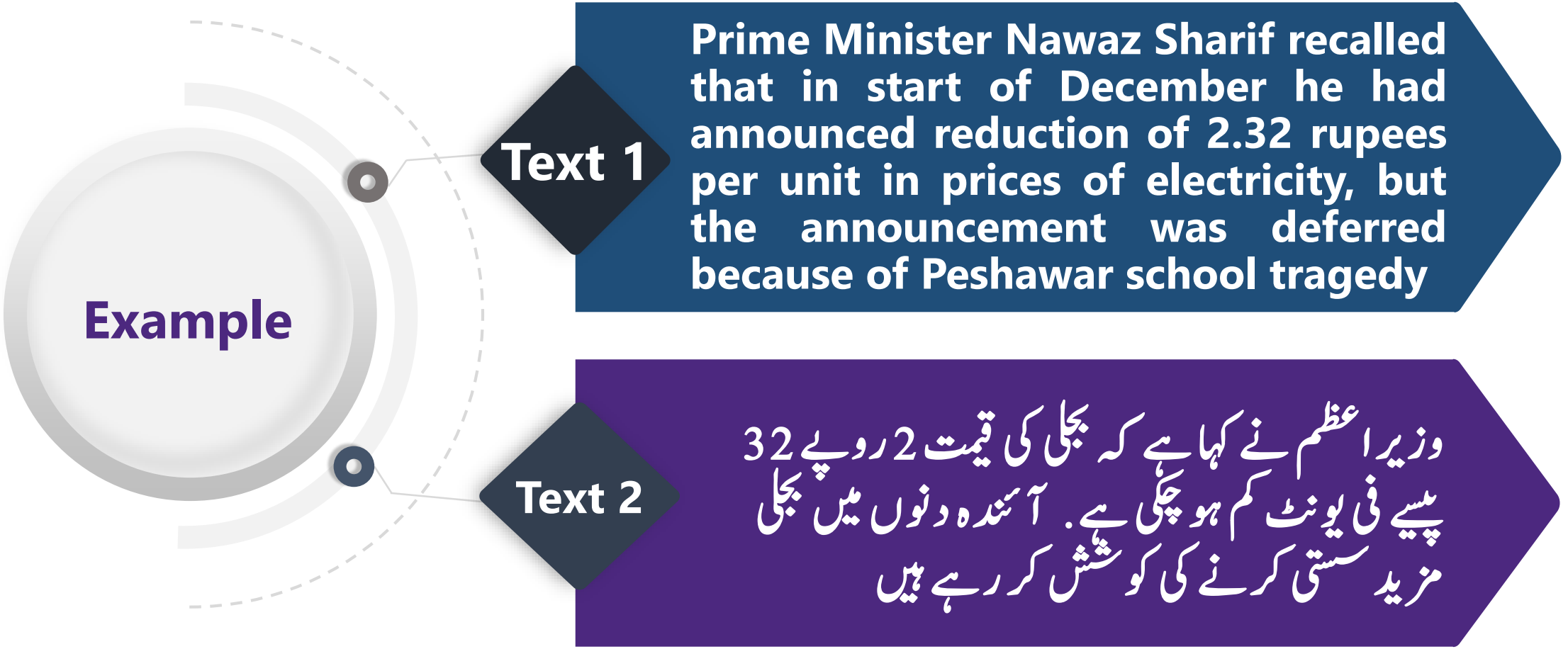
Example - Independently Written (Mono-lingual Settings)



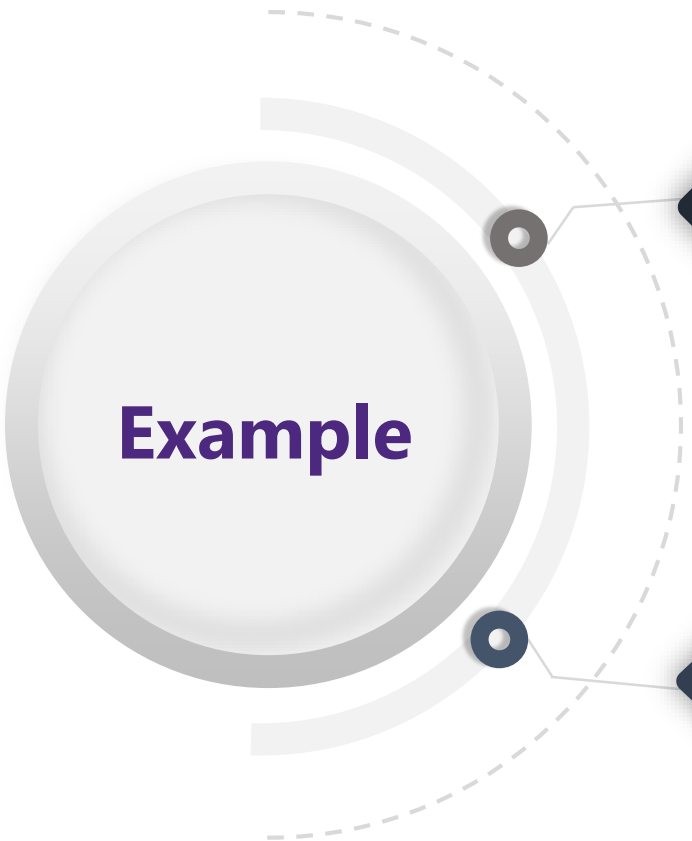
Example - Verbatim Copy / Exact Copy (Cross-lingual Settings)



Example - Paraphrased Copy (Cross-lingual Settings)



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 لیٹر تک کمی کا اعلان کر دیا. پٹرول 6.25, ڈیزل 7.86, مٹی کا تیل 11.26 روپے لٹر سستا

Your Turn

Write at least 12 examples for each of the following: 6 examples for mono-lingual text reuse and 6 examples for cross-lingual text reuse (2 Wholly Derived Examples, 2 Partially Derived Examples and 2 Non-Derived Examples)

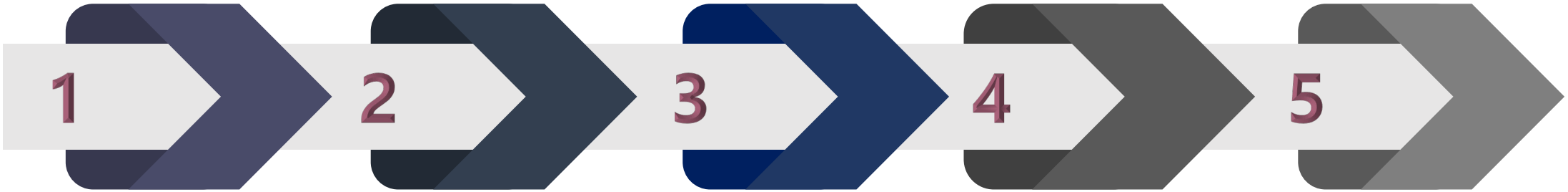
- Text Reuse at Word Level
- Text Reuse at Phrasal Level
- Text Reuse at Sentence Level
- Text Reuse at Passage / Paragraph Level
- Text Reuse at Document Level

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

Summary - Basics of Text Reuse (Cont.)

Text Reuse may occur at five levels:



1. **Word Level**
2. **Phrasal Level**
3. **Sentence Level**
4. **Passage / Paragraph Level**
5. **Document Level**

Summary - Basics of Text Reuse

- 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

It’s Poetry Time

Importance of Poetry

جوانی دیوانی کا دف بیوی سے مرتا ہے

بیوی کا دف کثرتِ اولاد سے

اولاد کا دف سائنس سے

اور سائنس کا دف شاعری سے مرتا ہے

مشتاق احمد یوسفی

دف مارنا -- کسی چیز کی تیزی کا کم ہونا

شعر

عقل والوں کے نصیبوں میں کہاذوقِ جنوں
عشق والے ہیں جو ہر چیز لٹادیتے ہیں

Stay Motivated

Importance of Motivation

گاڑی باہر کے دھکوں سے نہیں اندر کی طاقت سے چلتی ہے.

گاڑی ایک کروڑ کی ہو اور پٹرول نہ ہو تو نہیں چلے گی.

Motivation is the Fuel of Life

Tips - To Stay Motivated and become a Great Researcher

- **Make a Schedule of 24 Hours and Live a Live a Balance Life**
  - **Watch Seminar - Balanced Life is Ideal Life**
    - **Download Link:** *https://drive.google.com/open?id=1jet6r1QOtAB16Glpgiq18EeXaCkeVJUp*
- **On Daily Basis, read and enjoy**
  - **Poetry**
  - **Jokes**

Motivational Seminar

Title: Guru ki har bat gur hoti hai Gur ko nhe Guru ko pakar

Speaker: Dr. Rao Muhammad Adeel Nawab

Downlaod Link: Video + Slides

https://drive.google.com/open?id=1SQ5yUroaPT5dRcY-8K7IVC5SQrQy8Lsh

Task

Summarize main points of the motivational seminar

Basics of Plagiarism

Plagiarism

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)**

Plagiarism (Cont.)

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

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

Levels of Plagiarism

Verbatim

The original text is reused as verbatim (word to word copy) or with minor modifications to create the plagiarized document

Levels of Plagiarism (Cont.)

Paraphrased Plagiarism

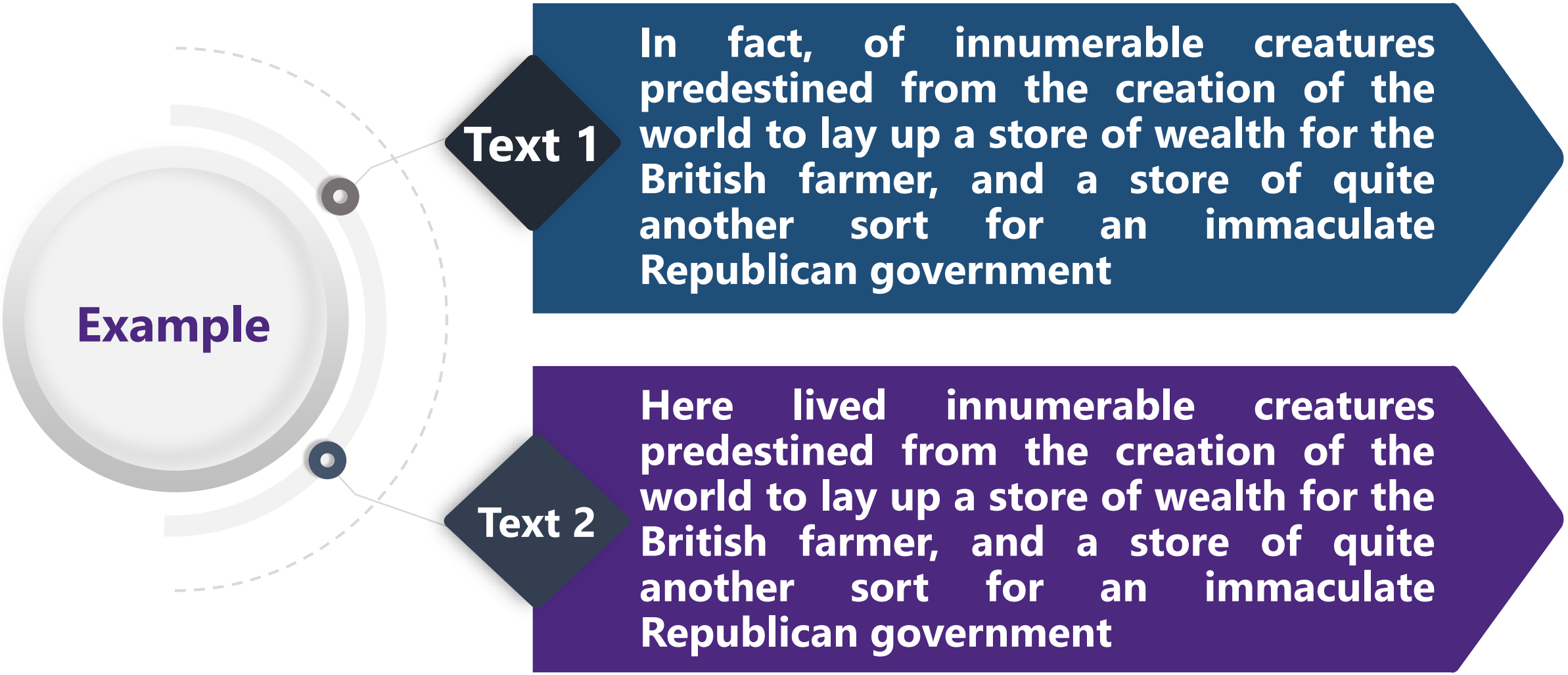
- **The original text is heavily altered (or paraphrased) to create the plagiarized document**
- **Paraphrasing can be as:**
  - **Light Revision**
    - **Source text is slightly paraphrased**
  - **Heavy Revision**
    - **Source text is heavily paraphrased**

Levels of Plagiarism (Cont.)

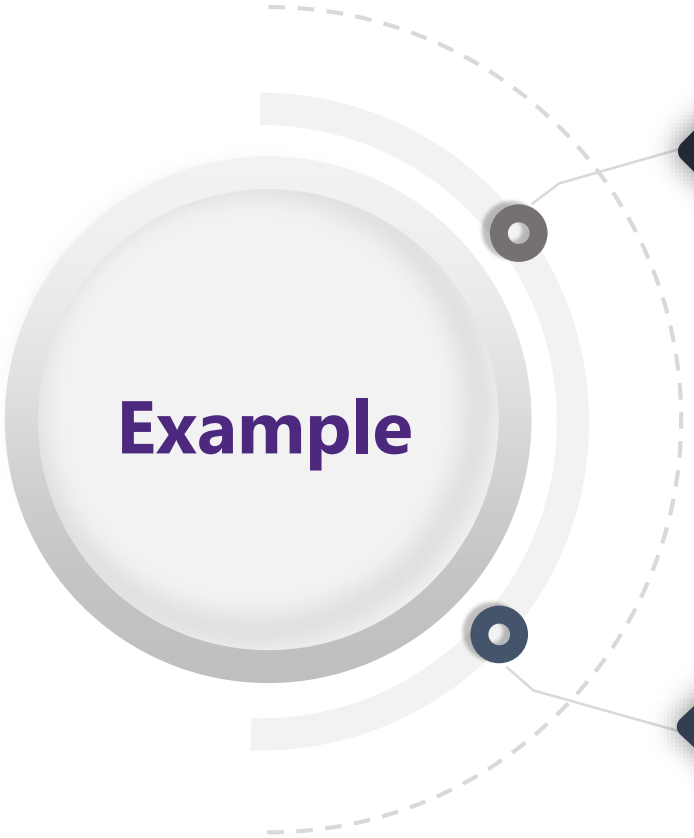
Plagiarism of Idea

- **The idea of the original text is reused without dependence on the words or form of the source**

Example - Verbatim Plagiarism



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)

Example - Plagiarism of Idea

Text 1

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.

Plagiarism Detection – Task

Given

- A suspicious text (input)

Identify

- The source(s) of plagiarism

Plagiarism Detection – Input and Output

Input

- Suspicious Text

Output

- Plagiarized / Non-Plagiarized

Plagiarism Detection - Two Levels of Rewrite

Plagiarized

- When any type of plagiarism is occurred between documents they were called plagiarized

Non-Plagiarized

- When no type of plagiarism is occurred between documents they were called non plagiarized

Plagiarism Detection - Four Levels of Rewrite

The Plagiarized cases can be further categorized into three categories

1 - Near Copy

- When suspicious text is created by simply copying and pasting text from source document(s)

2 - Light Revision

- When suspicious text is created by applying small modification like synonyms replacement and altering grammatical structure

Plagiarism Detection - Four Levels of Rewrite (Cont.)

3 - Heavy Revision

- When suspicious text is created by rephrasing the text to generate the meaning
  - It may include breaking source sentence into more than one sentences, merging two or more sentences into one, replacing words with appropriate synonyms or phrases, changing voice, changing tense etc.

4 - Non-Plagiarized

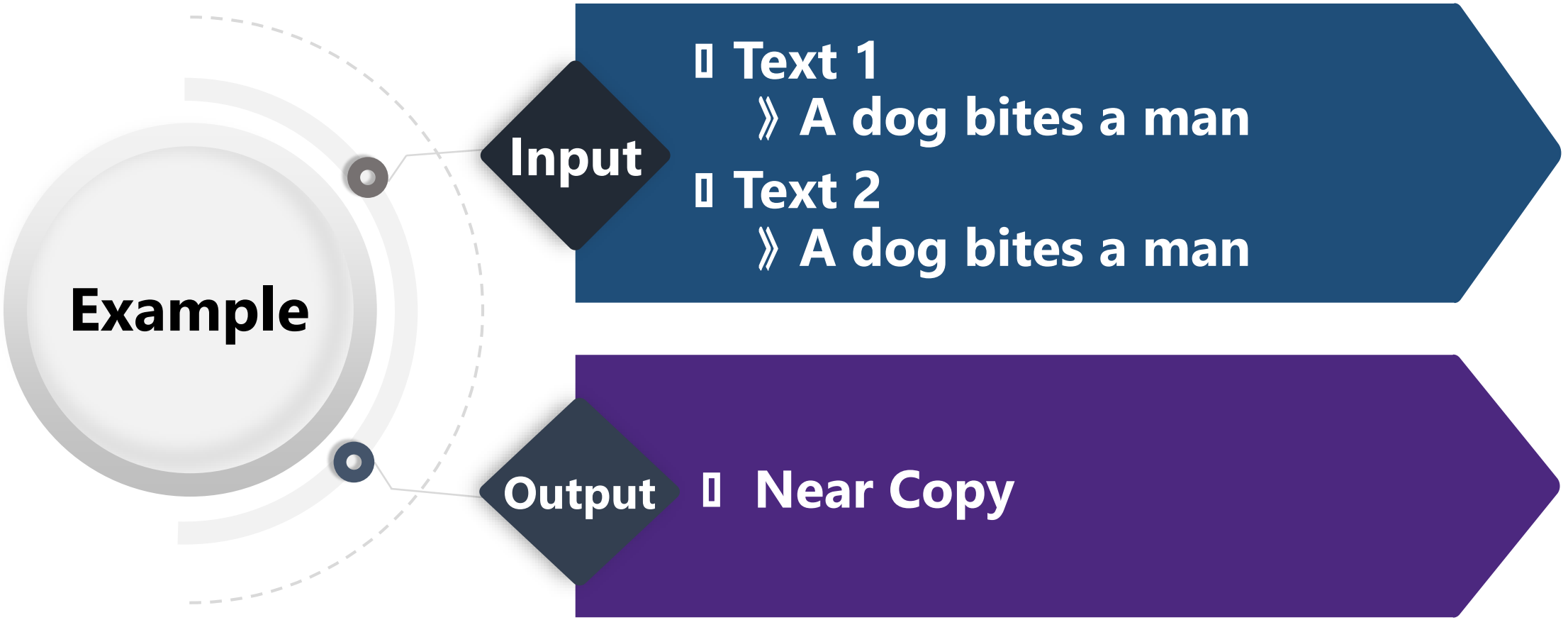
- When suspicious text is written independently

Important Note

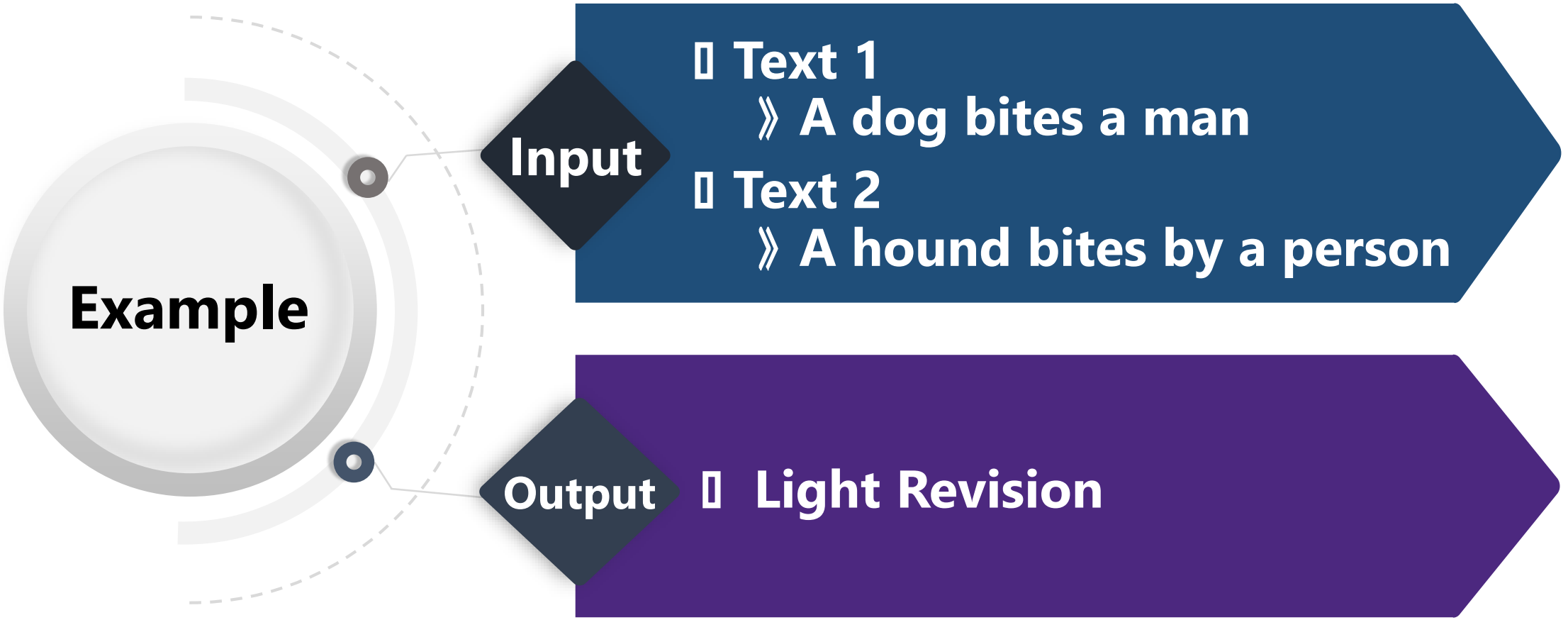
Documents that are independently written on the same topic are expected to have

Around 50% content overlap

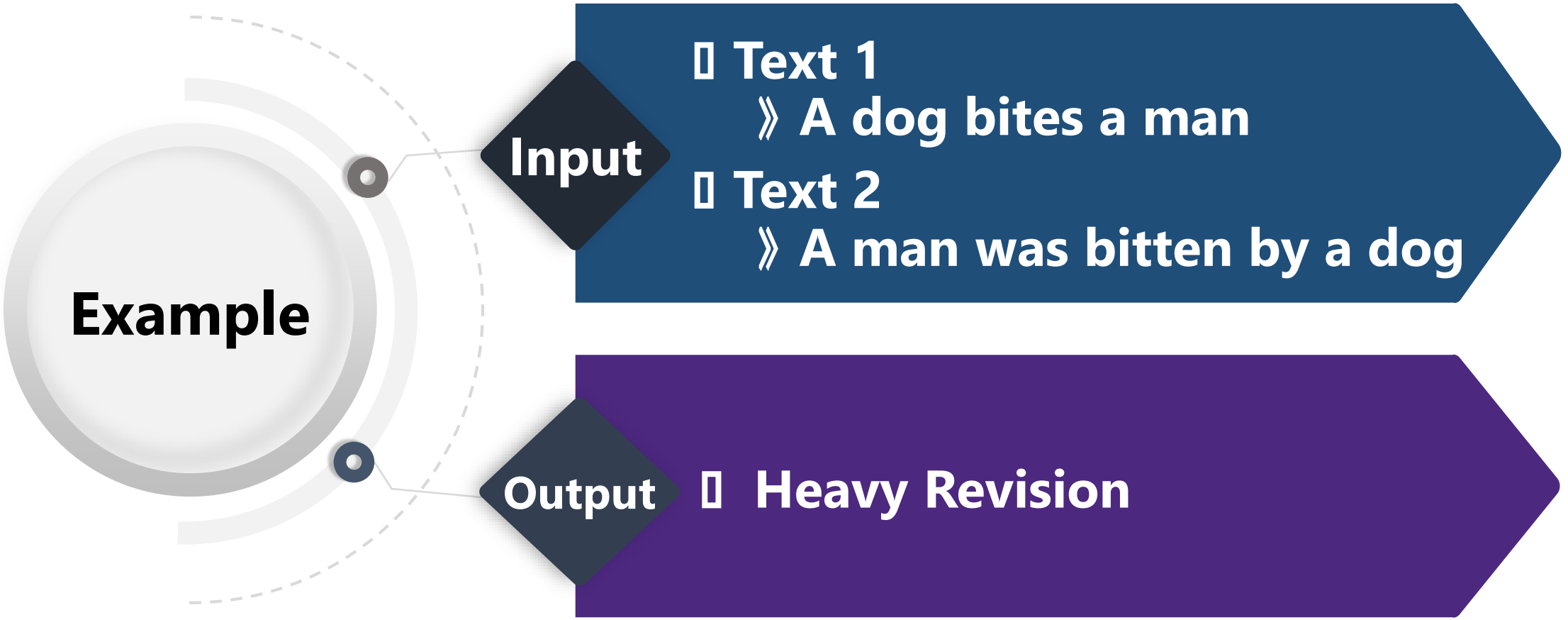
Example - Near Copy



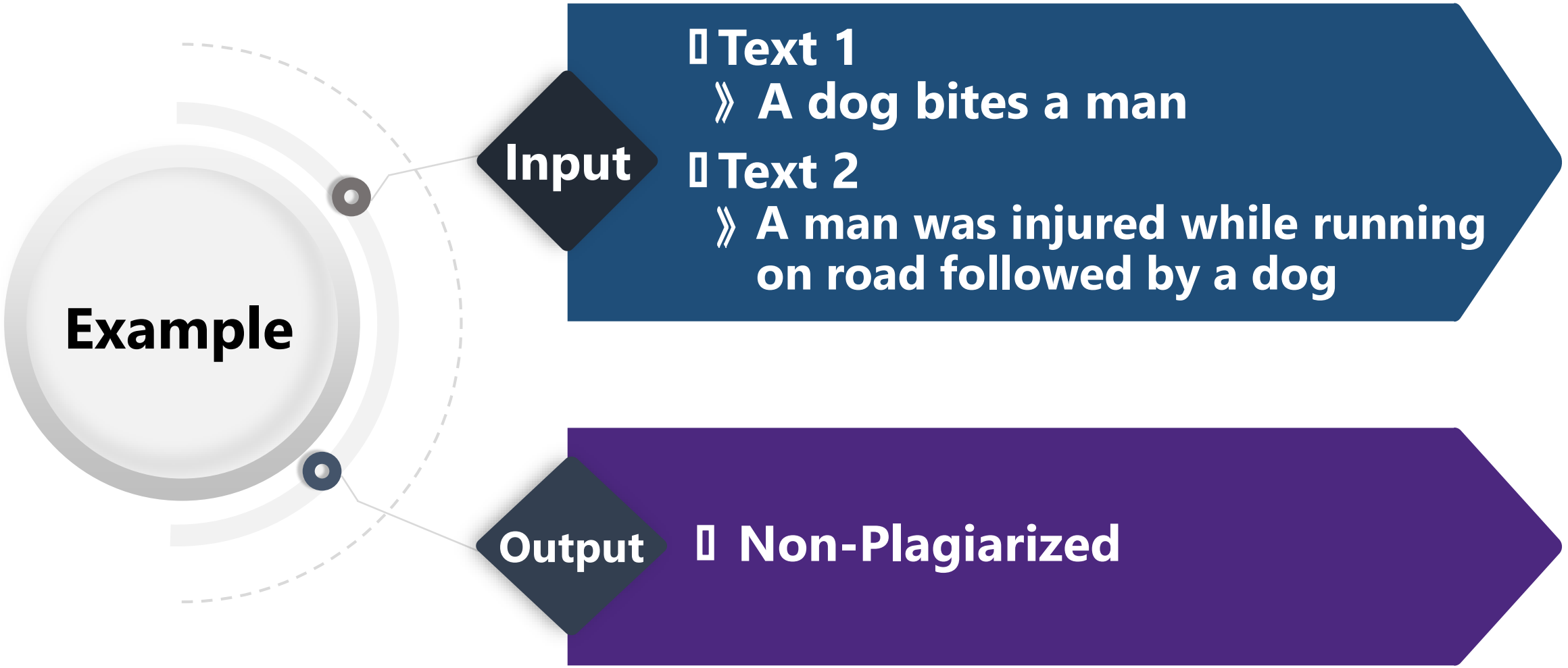
Example - Light Revision



Example - Heavy Revision



Example - Non-Plagiarized



Types of Plagiarism Cases

There are three main types of plagiarism cases

1 Artificial

> **Artificial cases of plagiarism are generated by using Automatic Text Altering tools to obfuscate the source text for plagiarism**

Types of Plagiarism Cases (Cont.)

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

Type of Plagiarism Cases (Cont.)

2 Simulated / Manual

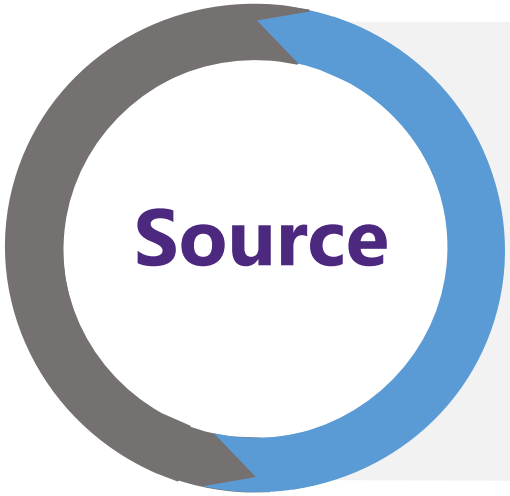
The original text is paraphrased by humans to create the cases of plagiarism

3 Real

Real cases of plagiarism are those which occurred in the real world

- For example, Karl-Theodor zu Guttenberg (German Defense Minister) PhD thesis proved plagiarized
- URL: https://www.theguardian.com/world/2011/mar/01/german-defence-minister-resigns-plagiarism

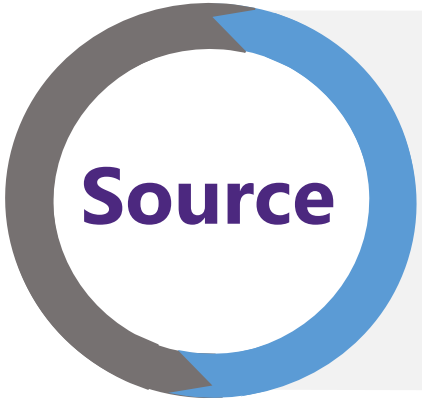
Example – Artificial (None Obfuscation)



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

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

Example – Manual (Simulated Obfuscation)



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

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

Real

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

Types of Plagiarism Detection

Two main types of plagiarism detection

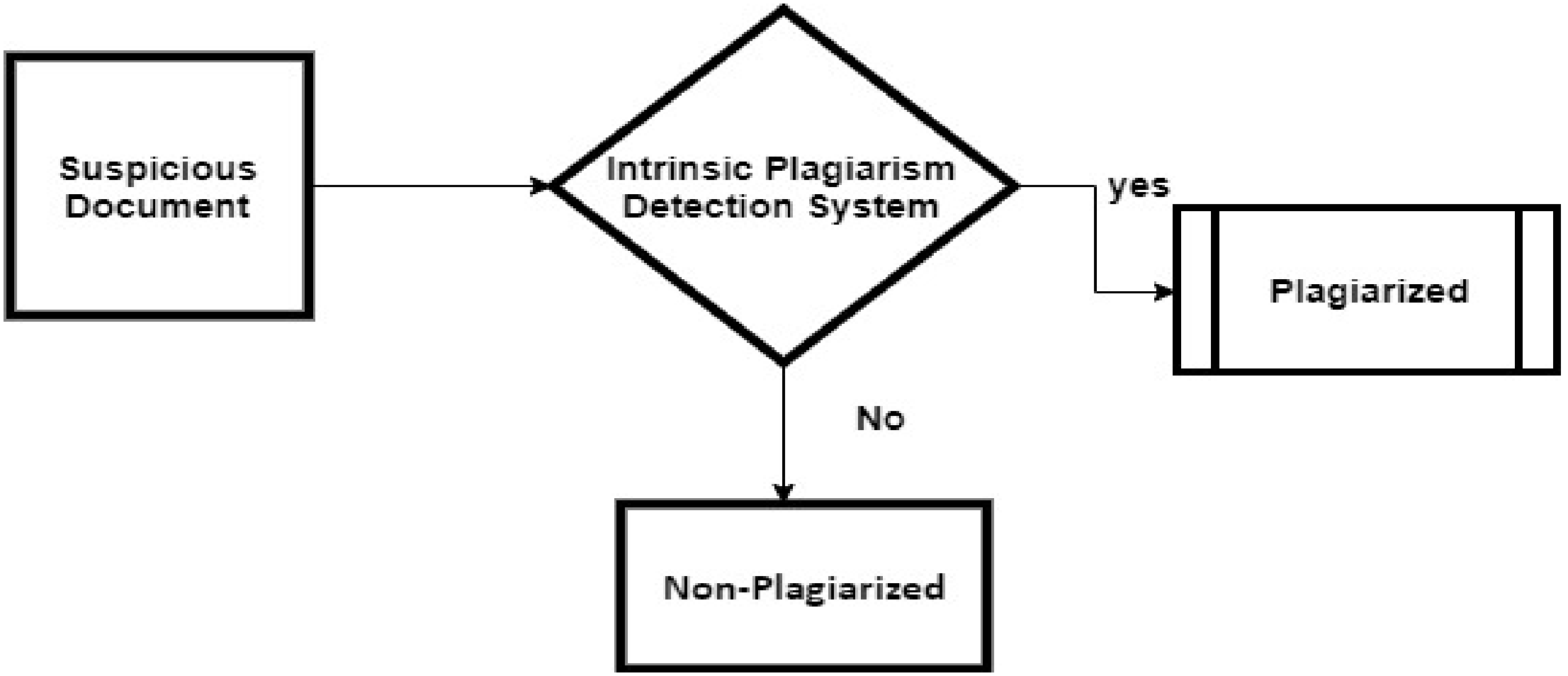
Intrinsic Plagiarism Detection

- 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).

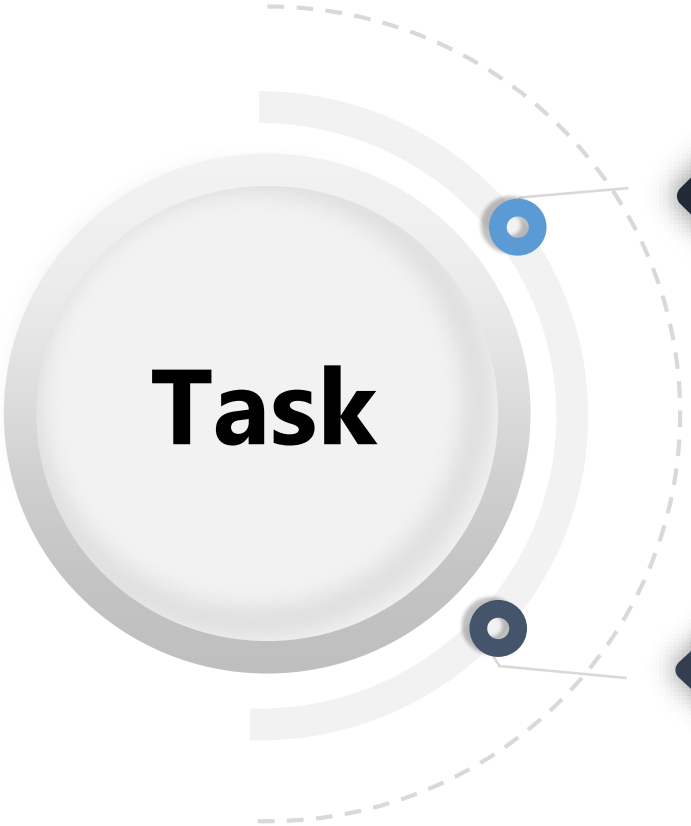
Extrinsic Plagiarism Detection

- 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

Intrinsic Plagiarism Detection – Task



Intrinsic Plagiarism Detection – Task



Given

- A suspicious document (input)

Identify

- Portion(s) of text whose writing style is significantly different form the remaining text (output)

Intrinsic Plagiarism Detection – Input and Output

Input | **A Suspicious Text**

Output | **Portion(s) of text whose writing style is significantly different form the remaining text**

Note | **If whose writing style is one or more portion(s) of text is significantly different form the remaining text then the suspicious document is plagiarized otherwise non-plagiarized**

Example – Intrinsic Plagiarism Detection

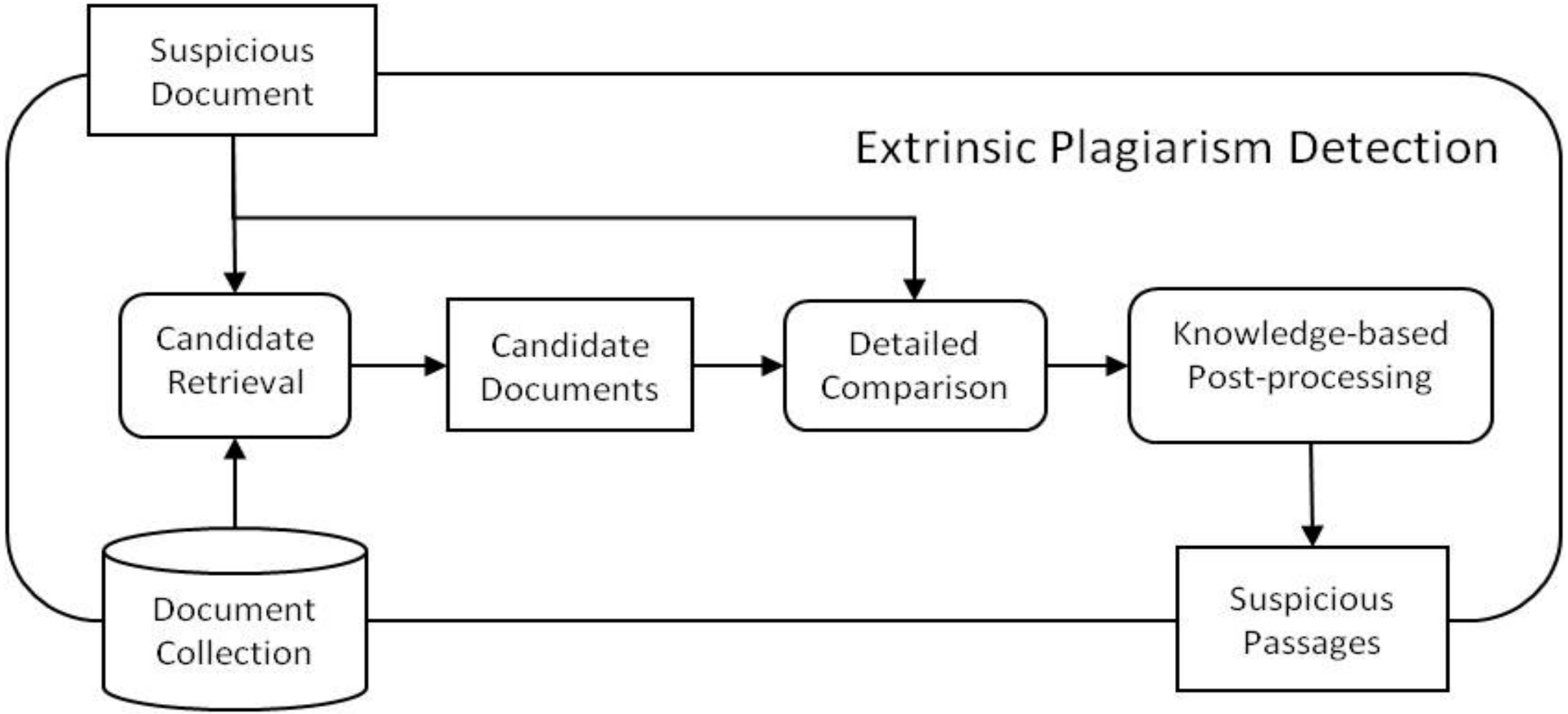
Given (Suspicious Document)

- 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 institution of the world i.e. MIT, U.S.A. He is humble and nice. Rasheed always try to help others

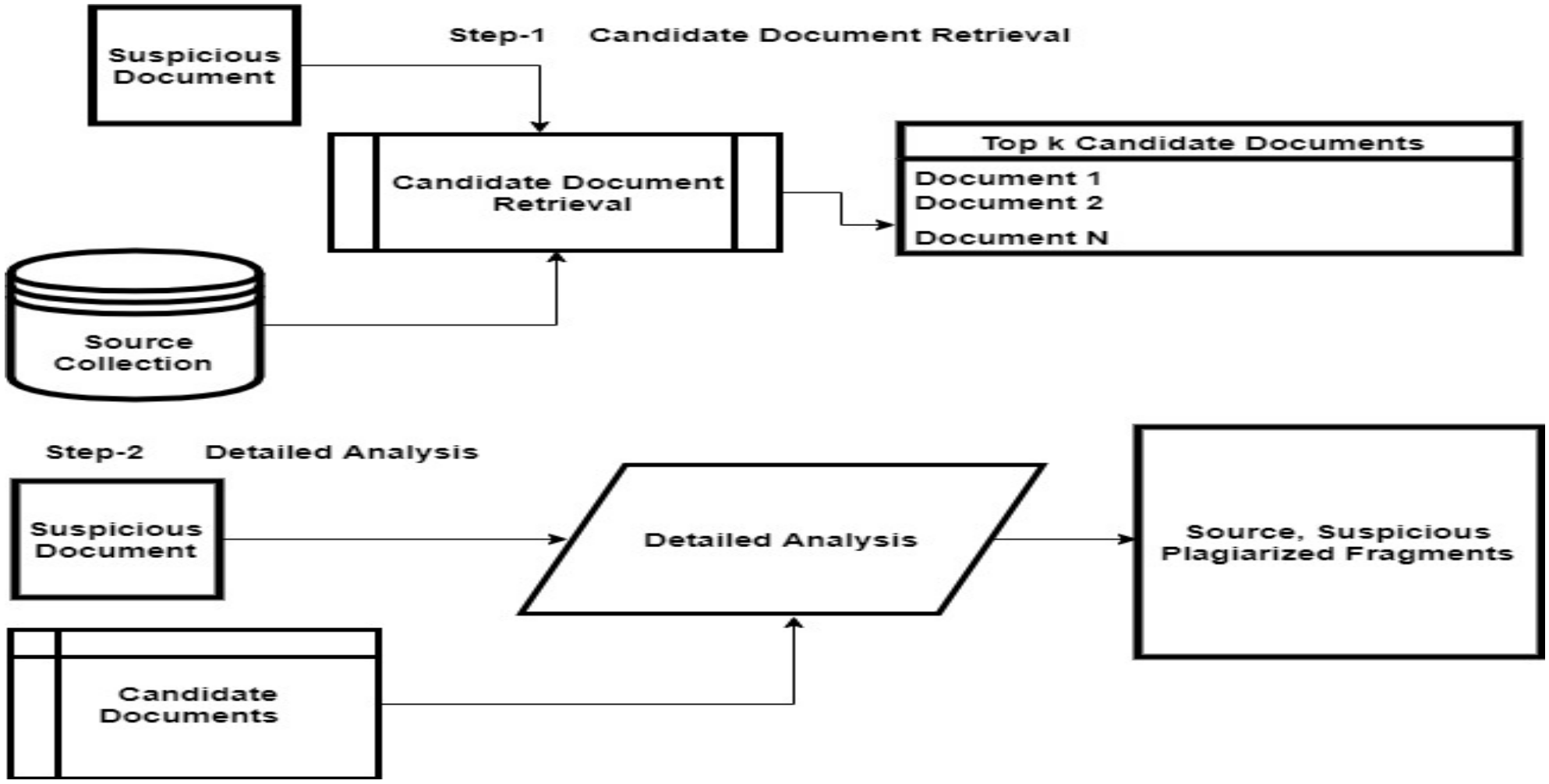
Output

- 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 institution of the world i.e. MIT, U.S.A.

Extrinsic Plagiarism Detection - Task



Extrinsic Plagiarism Detection – Task (Cont.)



Example – Extrinsic Plagiarism Detection

Given

Suspicious Collection

- Containing two suspicious documents

Source Collection

- Containing five source documents

Example – Extrinsic Plagiarism Detection (Cont.)

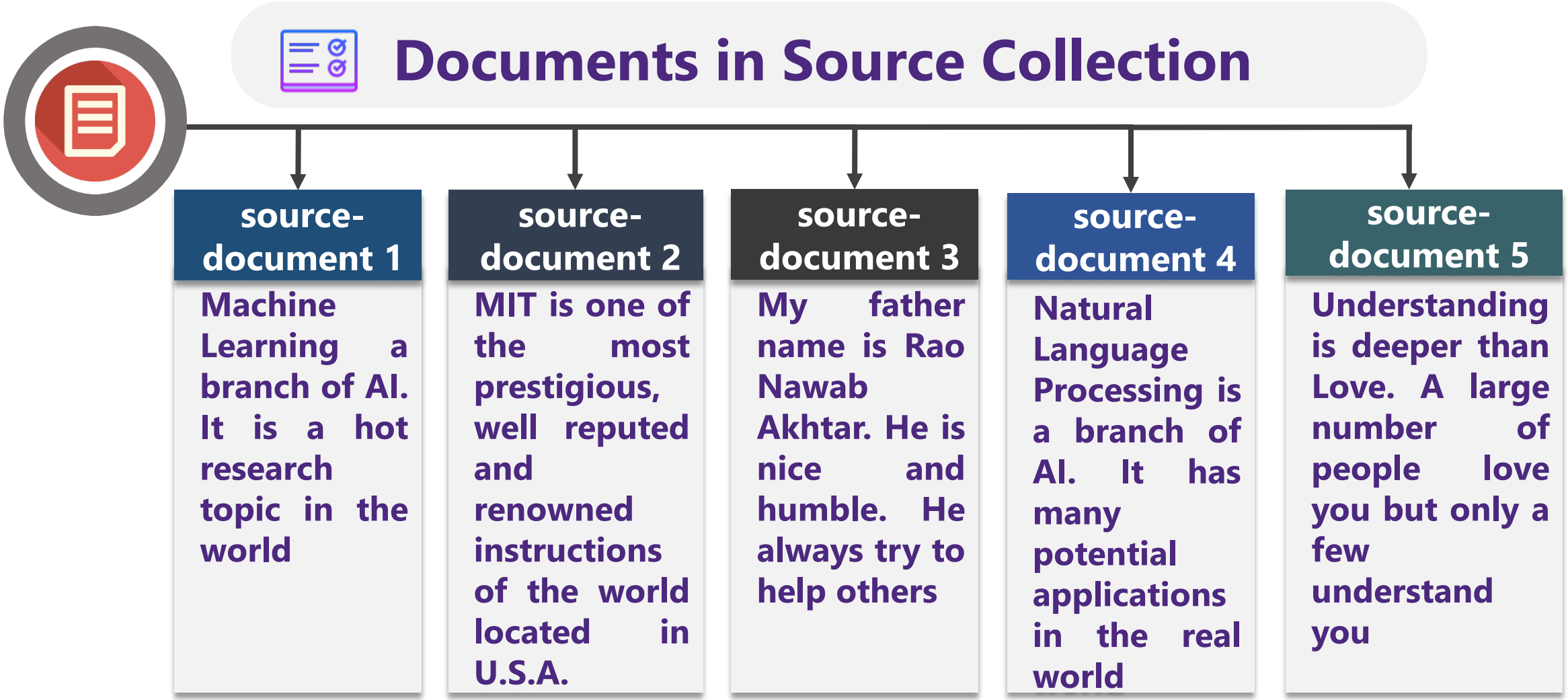
Documents in Suspicious Collection

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.

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.

Example – Extrinsic Plagiarism Detection (Cont.)



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

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

Example – Extrinsic Plagiarism Detection (Cont.)

Given

- **Query**
  - **Two Separate Queries**
    - **suspicious-document-01**
    - **suspicious-document-02**
- **Static Collection of Documents**
  - **Source Collection (containing 5 documents)**

Example – Extrinsic Plagiarism Detection (Cont.)

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-02 - Potential Candidate Source Document(s)
  - source-document-01
  - source-document-04
  - source-document-05

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

Example – Extrinsic Plagiarism Detection (Cont.)

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**

Example – Extrinsic Plagiarism Detection (Cont.)

Output – Detailed Analysis

- **Output - suspicious-document-01**
  - **suspicious-document-01 is Plagiarized**
  - **Source(s) of Plagiarism**
    - **suspicious-document-02**
    - **suspicious-document-03**

Example – Extrinsic Plagiarism Detection (Cont.)

Suspicious-Source Fragment Pairs

1 **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.

Example – Extrinsic Plagiarism Detection (Cont.)

2 **Fragment Pair 2 - suspicious-document-01, source-document-03**

- **suspicious-document-01**
  - » **He is humble and nice. Rasheed always try to help others.**
- **Source-document-03**
  - » **He is nice and humble. He always tries to help others.**

Example – Extrinsic Plagiarism Detection (Cont.)

Output – Detailed Analysis

- **Output - suspicious-document-02**
  - **suspicious-document-01 is Non - Plagiarized**

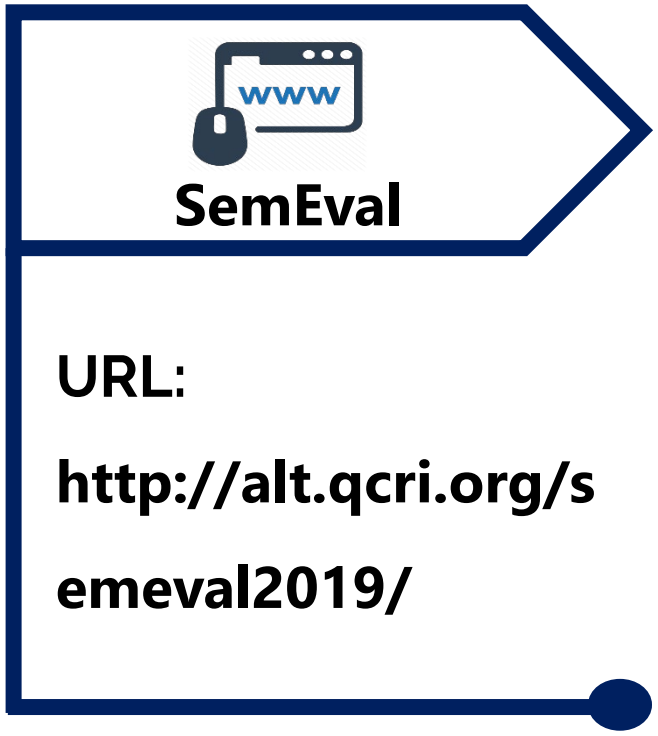
Shared Tasks on Text Reuse and Plagiarism

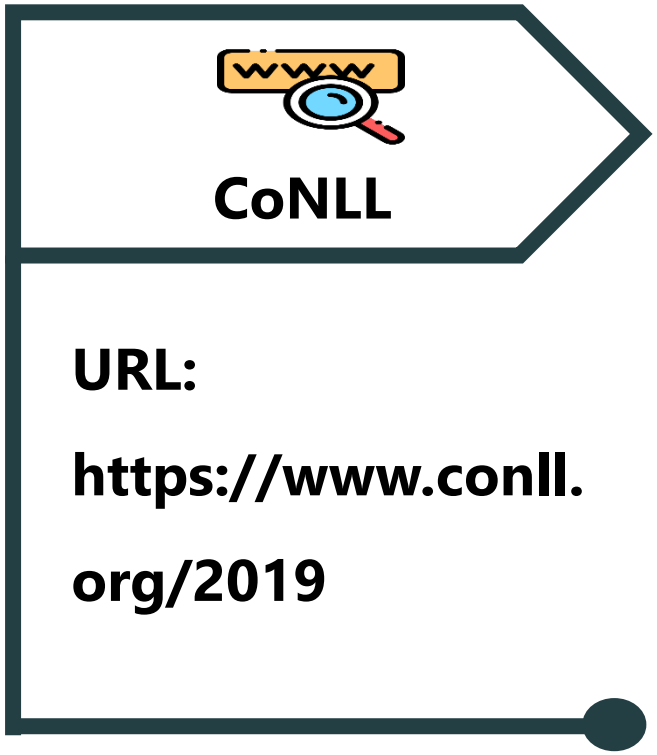
PAN

PAN is a series of scientific events and shared tasks on digital text forensics and stylometry

https://pan.webis.de/

Main Shared Tasks on Natural Language Processing





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

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

Summary - Basics of Plagiarism (Cont.)

- 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,

Summary - Basics of Plagiarism (Cont.)

(2) Simulated / Manual Cases of Plagiarism - the original text is paraphrased by humans to create the cases of plagiarism and **(3) Real Cases of Plagiarism** - 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

It’s Poetry Time

Importance of Poetry

جوانی دیوانی کا دف بیوی سے مرتا ہے

بیوی کا دف کثرتِ اولاد سے

اولاد کا دف سائنس سے

اور سائنس کا دف شاعری سے مرتا ہے

مشتاق احمد یوسفی

دف مارنا -- کسی چیز کی تیزی کا کم ہونا

شعر

آئینہ بنتا ہے رگڑے لاکھ جب کھاتا ہے دل
کچھ نہ پوچھو دل بڑی مشکل سے بن پاتا ہے دل

Stay Motivated

Importance of Motivation

گاڑی باہر کے دھکوں سے نہیں اندر کی طاقت سے چلتی ہے.

گاڑی ایک کروڑ کی ہو اور پٹرول نہ ہو تو نہیں چلے گی.

Motivation is the Fuel of Life

Tips - To Stay Motivated and become a Great Researcher

- **Make a Schedule of 24 Hours and Live a Live a Balance Life**
  - **Watch Seminar - Balanced Life is Ideal Life**
    - **Download Link:** *https://drive.google.com/open?id=1jet6r1QOtAB16Glpgiq18EeXaCkeVJUp*
- **On Daily Basis, read and enjoy**
  - **Poetry**
  - **Jokes**

Motivational Seminar

Title: Kamal Kiya Hai - Hussan Ho Aur Nazakat Na Ho

Speaker: Dr. Rao Muhammad Adeel Nawab

Downlaod Link: Video + Slides

https://drive.google.com/open?id=1RZOgjND7LiQSe0onFW3fo-GHjZsieTVA

Task

Summarize main points of the motivational seminar

Data Annotation for Text Reuse Detection

Data vs Information

PAN

- Raw Facts and Figures
- Varieties of Data

Varieties Of Data

Structured Data

Data is stored, processed, and manipulated in a traditional Relational Database Management System (RDBMS)

Unstructured Data

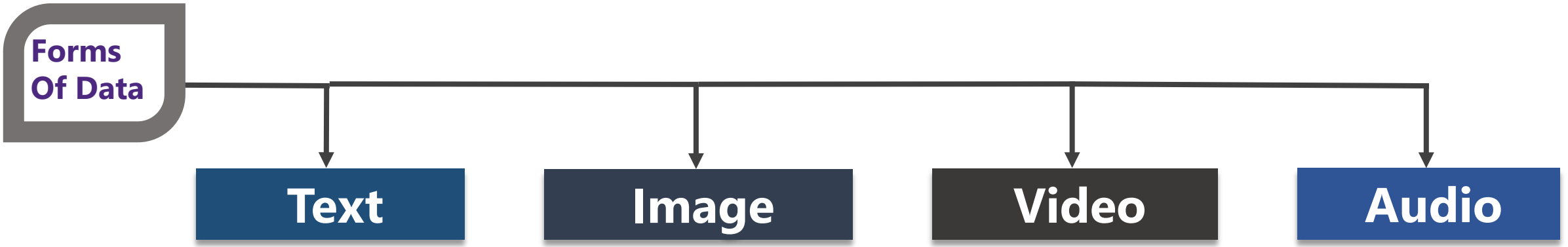
Data that is commonly generated from human activities and doesn't fit into a structured database format

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

Data vs Information (Cont.)

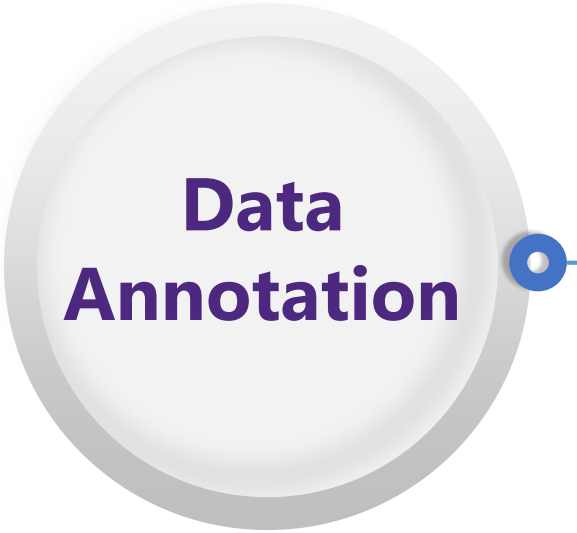
Main Forms Of Data



INFORMATION

Processed form of Data

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	
Comment / Review	**Sentiment**
iPhone7 is a good mobile	Positive
Battery of this phone is bad	Negative
I am using iphone7	Neutral

Data Annotation – Gender Identification	
Comment / Review	**Gender**
iPhone7 is a good mobile	Male
Battery of this phone is bad	Male
I am using iphone7	Female

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

Example – Data Annotation for Text Reuse Detection

Raw Data Collection

- **Two Sources of Data**
  - **01 News Agencies articles**
    - Associated Press of Pakistan (APP)
    - Independent News Agency (INP)
  - **02 Newspapers stories**
    - Express Newspaper
    - Jang Newspaper
    - Nawa e Waqt

Note Multiple Newspapers can reuse text from one News Agency to produce their news stories

Example – Data Annotation for Text Reuse Detection (Cont.)

Raw Data Collection

- **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

Example – Data Annotation for Text Reuse Detection (Cont.)

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

Example – Extrinsic Plagiarism Detection (Cont.)

Annotation Process

Annotation Guidelines

- Read text pair
- Assign the label (from pre-defined set of labels) which maps to the most dominating label

Example – Extrinsic Plagiarism Detection (Cont.)

- **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

Example – Data Annotation for Text Reuse Detection (Cont.)

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

Example – Data Annotation for Text Reuse Detection (Cont.)

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

Example – Data Annotation for Text Reuse Detection (Cont.)

Annotation Process

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

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	Partially Derived
This is my country	I live in Lahore, Pakistan	Non - Derived
I like your car	I like your car	Wholly Derived

Example – Data Annotation for Text Reuse Detection (Cont.)

Annotation Process

01 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

Inter-Annotator Agreement

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

Example – Data Annotation for Text Reuse Detection (Cont.)

02 IAA is computed to drive 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

Example – Data Annotation for Text Reuse Detection (Cont.)

Annotation Process

- Combine Annotations of A & B to Compute Inter Annotator Agreement (IAA)

$$\text{Inter-Annotator Agreement} = \frac{4}{6} = 0.667$$

Example – Data Annotation for Text Reuse Detection (Cont.)

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

Example – Data Annotation for Text Reuse Detection (Cont.)

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

Example – Corpus Standardization

Two Main Formats to Standardize Corpus

1. CSV
2. XML

Example – Corpus Standardization

Corpus Standardization in CSV Format

L4

	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			

Example – Corpus Standardization

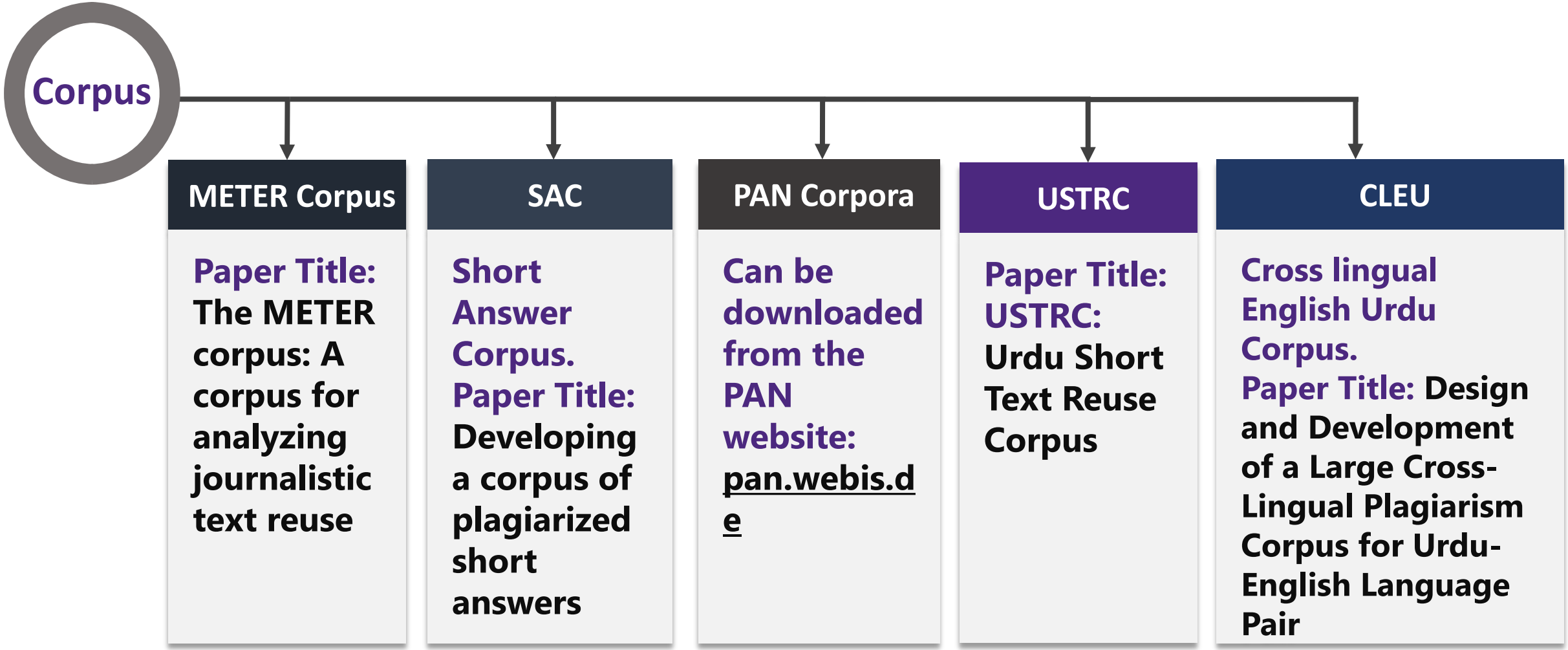
Corpus Standardization in XML Format

File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?

dice-score-similarityScore_mappedPOSS.csv | overlap-score-similarityScore_mappedPOSS.csv | toy.csv | toy-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>
```

Benchmark Corpora for Text Reuse and Plagiarism Detection



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

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

Summary - Data Annotation for Text Reuse Detection

Main Steps to Create Benchmark Annotated Dataset

01 Raw Data Collection

- Data Source(s)
- Cleaning of Data
- Pre-processing of Data

02 Simulated / Manual Cases of Plagiarism

- Preparation of Annotation Guidelines
- Annotations
- Computing Inter-Annotator Agreement

03 Corpus Standardization

It’s Poetry Time

Importance of Poetry

جوانی دیوانی کا دف بیوی سے مرتا ہے

بیوی کا دف کثرتِ اولاد سے

اولاد کا دف سائنس سے

اور سائنس کا دف شاعری سے مرتا ہے

مشتاق احمد یوسفی

دف مارنا -- کسی چیز کی تیزی کا کم ہونا

غزل

کبھی کتابوں میں پھول رکھنا کبھی درختوں پہ نام لکھنا
ہمیں بھی ہے یاد آج تک وہ نظر سے حرف سلام لکھنا
وہ چاند چہرے وہ بہکی باتیں سلگتے دن تھے مہکتی راتیں
وہ چھوٹے چھوٹے سے کاغذوں پر محبتوں کے پیام لکھنا
گلاب چہروں سے دل لگانا وہ چپکے چپکے نظر ملانا
وہ آرزوؤں کے خواب بننا وہ قصۂ نا تمام لکھنا

غزل...

مرے نگر کی حسیں فضاؤ کہیں جو ان کا نشان پاؤ
تو پوچھنا یہ کہاں بسے وہ کہاں ہے ان کا قیام لکھنا
کھلی فضاؤں میں سانس لینا عبث ہے اب تو گھٹن ہے ایسی
کہ چاروں جانب شجر کھڑے ہیں صلیب صورت تمام لکھنا
گئی رتوں میں حسنؔ ہمارا بس ایک ہی تو یہ مشغلہ تھا
کسی کے چہرے کو صبح کہنا کسی کی زلفوں کو شام لکھنا

حسن رضوی

Methods for Text Reuse and Plagiarism Detection

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

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

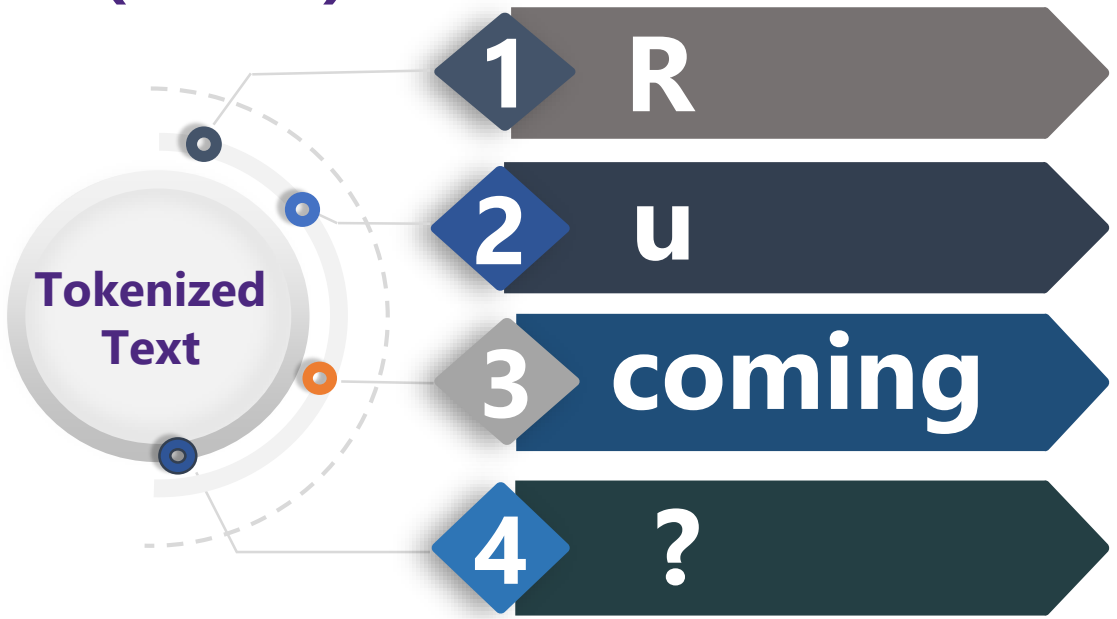
Definition

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**

Example – N-gram Generation from Input Text

- **Input Text**
  - R u coming?
- **Word Uni-grams (N = 1)**

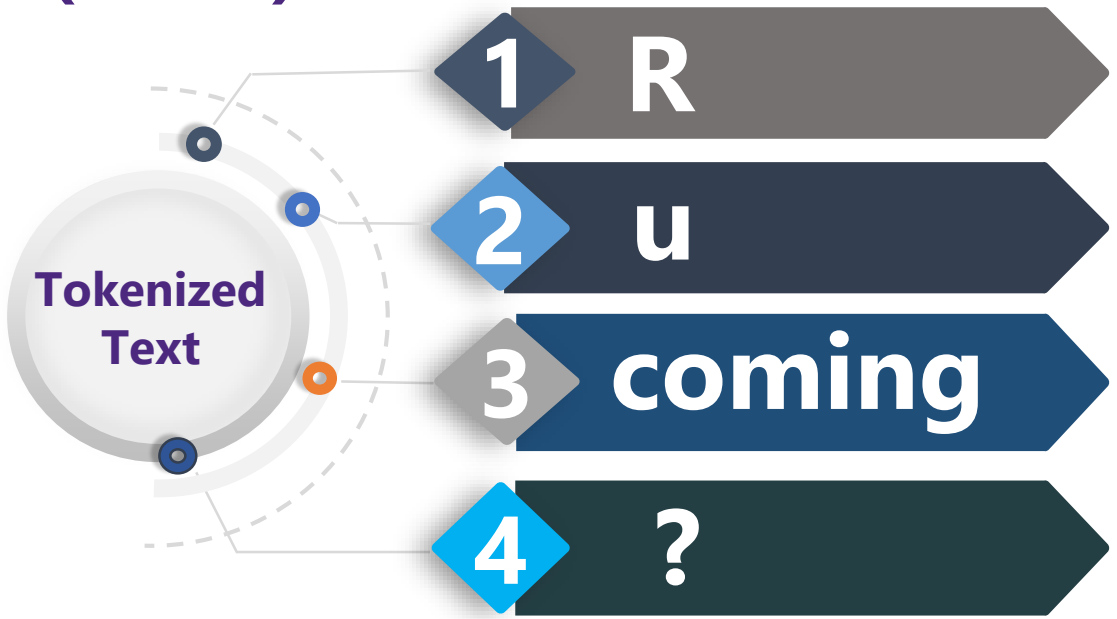


Set of Word Uni-grams

R	u	coming	?

Example – N-gram Generation from Input Text

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

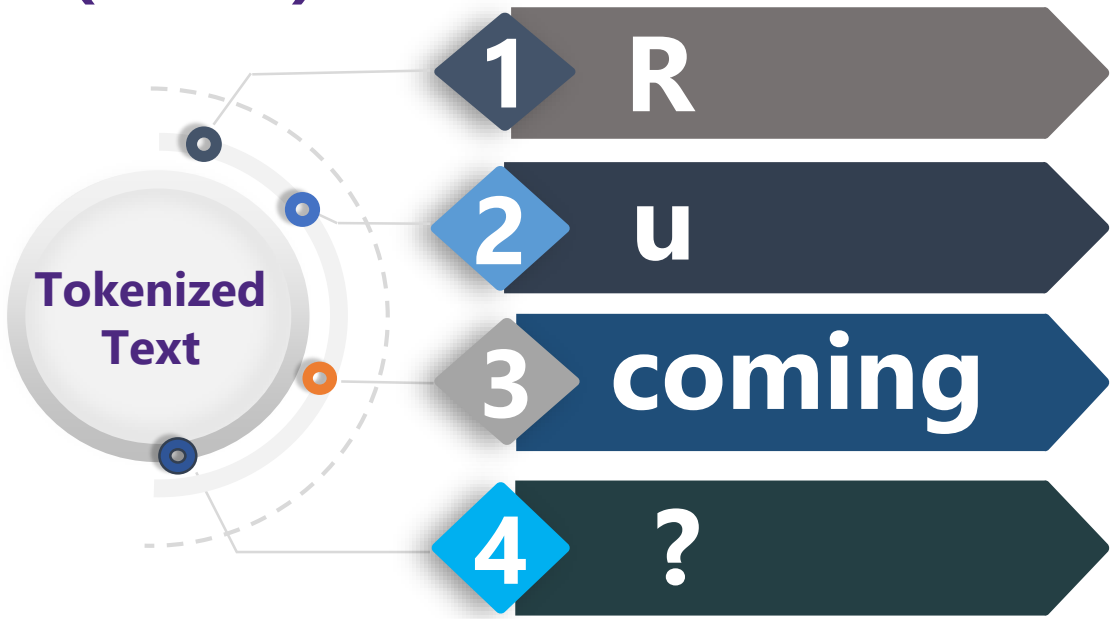


Set of Word Bi-grams

R u	U coming	Coming ?

Example – N-gram Generation from Input Text

- **Input Text**
  - R u coming?
- **Word Tri-grams (N = 3)**



Set of Word Tri-grams

R u coming	u coming ?

Example – N-gram Generation from Input Text

- **Input Text**
  - R u coming?
- **Character Tri-grams (N = 3)**
  - Tokenized Text:

r	u	coming?

  - Note that space is also a character

R		U		C	O	M	I	N	ing	?

  - Set of character Tri-grams

r u	u	co	com	omi	min	ing	gn?

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

Given

- A Text Pair (Text 1 and Text 2)

Steps - Commuting Similarity

1. Pre-process Input Text
   - Lower case
   - Punctuation Marks Removal
2. Convert Text 1 and Text 2 into Sets of N-grams

3 Compute Similarity Between Sets of N-grams using Overlap

Similarity Co-efficient i.e. Quantify the Degree of Similarity

Overlap Similarity

$$Overlap\ Similarity = \frac{|S(Text\ 1,n) \cap S(Text\ 2,n)|}{\min(|S(Text\ 1,n)|, |S(Text\ 2,n)|}$$

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

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

1 **Pre-process 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

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}

3 Compute Similarity Between Sets of N-grams using Overlap Similarity Co-efficient i.e. Quantify the Degree of Similarity

Overlap 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}|))))

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

Overlap Similarity

$$Overlap\ Similarity = \frac{|\{a, bites, a, man\}|}{\min(5,5)}$$

$$= \frac{2}{4} \quad = \mathbf{0.80}$$

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

1 **Pre-process 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

2 Convert Text 1 and Text 2 into Sets of N-grams

- Text 1 Word Bigrams (S (Text 1, 2)) = {a dog, dog bites, bites a, a man}
- Text 2 Word Bigrams (S (Text 1, 2)) = {a, hound, bites, a, man}

3 Compute Similarity Between Sets of N-grams using Overlap

Similarity Co-efficient i.e. Quantify the Degree of Similarity

Overlap 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}|)))

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

Overlap Similarity

$$Overlap\ Similarity = \frac{|\{\text{bites a}, a\ \text{man}\}|}{\min(4,4)}$$

$$= \frac{2}{4} \quad = 0.5$$

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

1 **Pre-process 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.

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

Text Pair After Pre-processing

- Text 1: a dog bites a man
- Text 2: a hound bites a man

2 Convert Text 1 and Text 2 into Sets of N-grams

- Text 1 Word Bigrams (S (Text 1, 3)) = {a dog bites, dog bites a, bites a man}
- Text 2 Word Bigrams (S (Text 1, 3)) = {a hound bites, hound bites a , bites a man}

3 Compute Similarity Between Sets of N-grams using Overlap

Similarity Co-efficient i.e. Quantify the Degree of Similarity

Overlap 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}|)))

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

Overlap Similarity

$$Overlap\ Similarity = \frac{|\{\text{bites a man}\}|}{\min(3,3)}$$

$$= \frac{1}{3} \qquad = 0.33$$

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 for Mono-lingual Text Reuse and Plagiarism Detection

Methods based on structure

- Stop-words based n-grams overlap

Methods based on style

- Type token ratio
- Token ratio
- Sentence ratio

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 for Cross-lingual Text Reuse and Plagiarism Detection

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 for Cross-lingual Text Reuse and Plagiarism Detection

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

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

$$\text{Similarity Score} = \frac{len(LCS)}{min\,(len(Text\ 1), len(Text\ 2))}$$

$$\text{Similarity Score} = \frac{4}{min\,(5,5)} = \frac{4}{min\,(5,5)} = 0.80$$

Your Turn

- is to take at least three text pairs and compute similarity score between then using the Longest Common Subsequence (LCS) Approach

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 Embedding's and (7) Methods based on Deep Learning

It’s Poetry Time

Importance of Poetry

جوانی دیوانی کا دف بیوی سے مرتا ہے

بیوی کا دف کثرتِ اولاد سے

اولاد کا دف سائنس سے

اور سائنس کا دف شاعری سے مرتا ہے

مشتاق احمد یوسفی

دف مارنا -- کسی چیز کی تیزی کا کم ہونا

شعر

یہ کہ رہی ہے تجھے چھو کے آنے والی ہوا
اداس میں ہی نہیں بے قرار تو بھی ہے

Stay Motivated

Importance of Motivation

گاڑی باہر کے دھکوں سے نہیں اندر کی طاقت سے چلتی ہے.

گاڑی ایک کروڑ کی ہو اور پٹرول نہ ہو تو نہیں چلے گی.

Motivation is the Fuel of Life

Tips - To Stay Motivated and become a Great Researcher

- **Make a Schedule of 24 Hours and Live a Live a Balance Life**
  - **Watch Seminar - Balanced Life is Ideal Life**
    - **Download Link:** *https://drive.google.com/open?id=1jet6r1QOtAB16Glpgiq18EeXaCkeVJUp*
- **On Daily Basis, read and enjoy**
  - **Poetry**
  - **Jokes**

Motivational Seminar

Title: Power Of Appreciation

Speaker: Dr. Rao Muhammad Adeel Nawab

Downlaod Link: Video + Slides

https://drive.google.com/open?id=1XnwhNjrjHPBSv0VHYuvgS5m4A7QwNDTf

Task

Summarize main points of the motivational seminar

Evaluation Measures

Evaluation Measures

Precision

Precision (P) of a text reuse detection system is the proportion of the predicted positive cases that were correct

Precision

$$P = \frac{TP}{TP+FP}$$

Evaluation Measures

Recall

Recall (R) of a text reuse detection system is defined as the proportion of positive cases that were correctly identified

Recall

$$R = \frac{TP}{TP+FN}$$

Evaluation Measures

F_1 measure

F_1 measure is a specific relationship (harmonic mean) between precision (P) and recall (R)

F_1 measure

$$F_1 = \frac{2 * P * R}{P + R}$$

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

It’s Poetry Time

Importance of Poetry

جوانی دیوانی کا دف بیوی سے مرتا ہے

بیوی کا دف کثرتِ اولاد سے

اولاد کا دف سائنس سے

اور سائنس کا دف شاعری سے مرتا ہے

مشتاق احمد یوسفی

دف مارنا -- کسی چیز کی تیزی کا کم ہونا

غزل

رنجش ہی سہی دل ہی دکھانے کے لیے آ
آ پھر سے مجھے چھوڑ کے جانے کے لیے آ

کچھ تو مرے پندار محبت کا بھرم رکھ
تو بھی تو کبھی مجھ کو منانے کے لیے آ

پہلے سے مراسم نہ سہی پھر بھی کبھی تو
رسم و رہ دنیا ہی نبھانے کے لیے آ

غزل...

کس کس کو بتائیں گے جدائی کا سبب ہم
تو مجھ سے خفا ہے تو زمانے کے لیے آ

اک عمر سے ہوں لذت گریہ سے بھی محروم
اے راحت جاں مجھ کو رلانے کے لیے آ

اب تک دل خوش فہم کو تجھ سے ہیں امیدیں
یہ آخری شمعیں بھی بجھانے کے لیے آ

احمد فراز

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

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

Treating the Problem of Text Reuse Detection as Machine Learning Problem

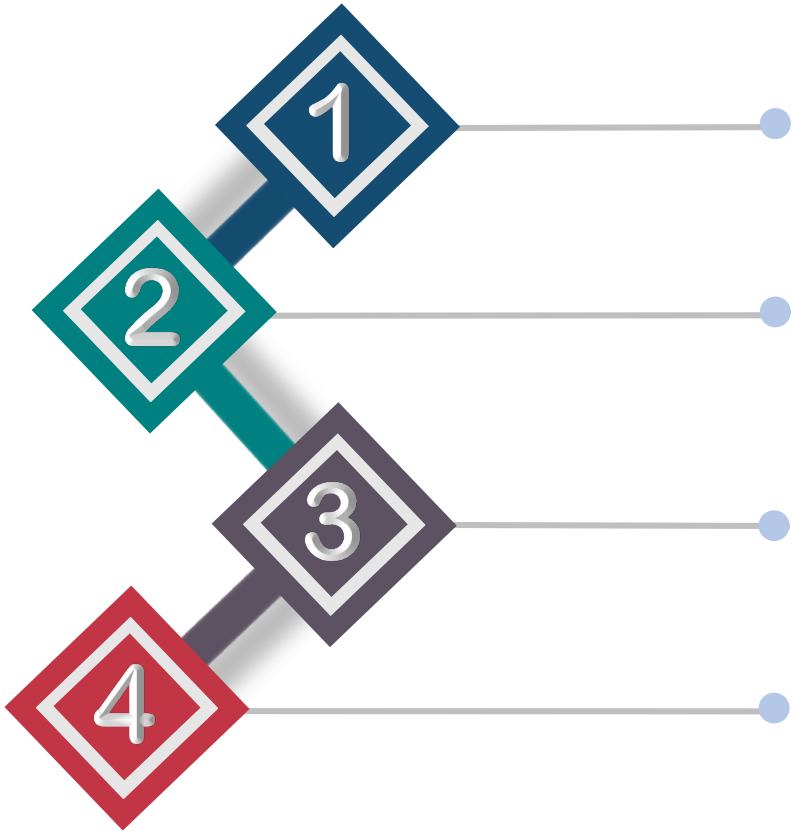
Question?	How to Transform Text Reuse Detection Problem to Supervised Text Classification Task?

For Supervised Text Classification Task

- Output must be associated with the input i.e. dataset must be annotated

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

Main Steps to Treat Text Reuse Detection as Supervised Text Classification Task



Pre-process input (i.e. text pair)

Feature Extraction from Input

Represent extracted features and output / label into a format that Machine Learning algorithms can understand (normally saved as CSV file)

Use features extracted in Step 2 to train and test Machine Learning algorithms

Experimental Setup

Problem of Text Reuse Detection is Treated as a Supervised Text Classification Task

- Two Versions of Classification

Binary Classification	Ternary Classification
Discriminate between two classes i.e. Derived vs Non-derived	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	Output
Text Pair	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

Experimental Setup

Technique (for Feature Extraction)

N-gram Overlap

- N-grams are word based
- N = 1 - 5

Evaluation Measure

- Precision
- Recall
- F_1

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)**

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

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 an 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.

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

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

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

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

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

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)

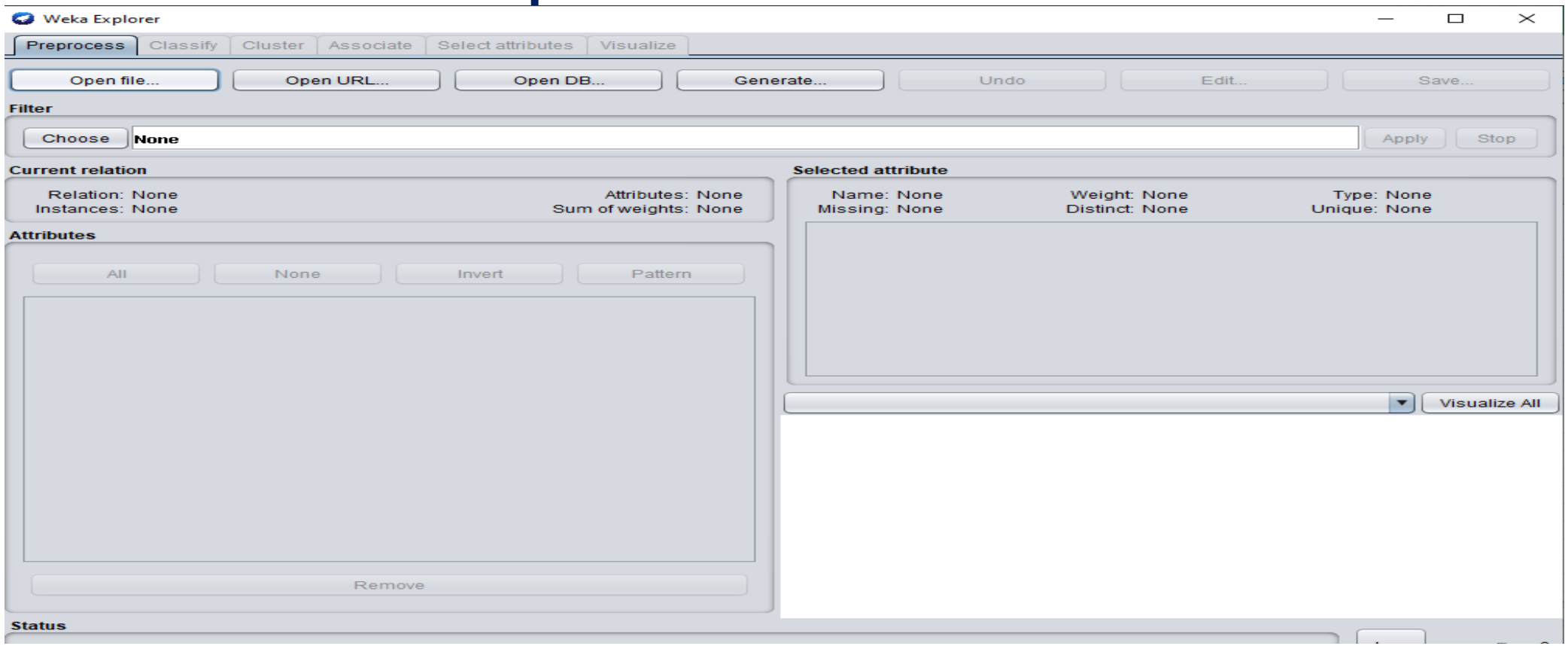
Uni-gram-Scores	Bi-gram-Score	Tri-gram-score	Four-gram-score	Five-gram-Score	Label
0.6	0.5	0.33	0	0	WD
0.6	0.25	0	0	0	PD
0.4	0	0	0	0	ND
0.5	0.33	0	0	0	ND
0	0	0	0	0	PD
1	1	1	1	0	WD
1	1	1	1	1	WD
0.33	0	0	0	0	PD
0.67	0	0	0	0	ND
0.17	0	0	0	0	ND
0.12	0	0	0	0	PD
0	0	0	0	0	PD
0.75	0.57	0.33	0	0	WD
0.57	0.31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

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

Ternary Classification using WEKA

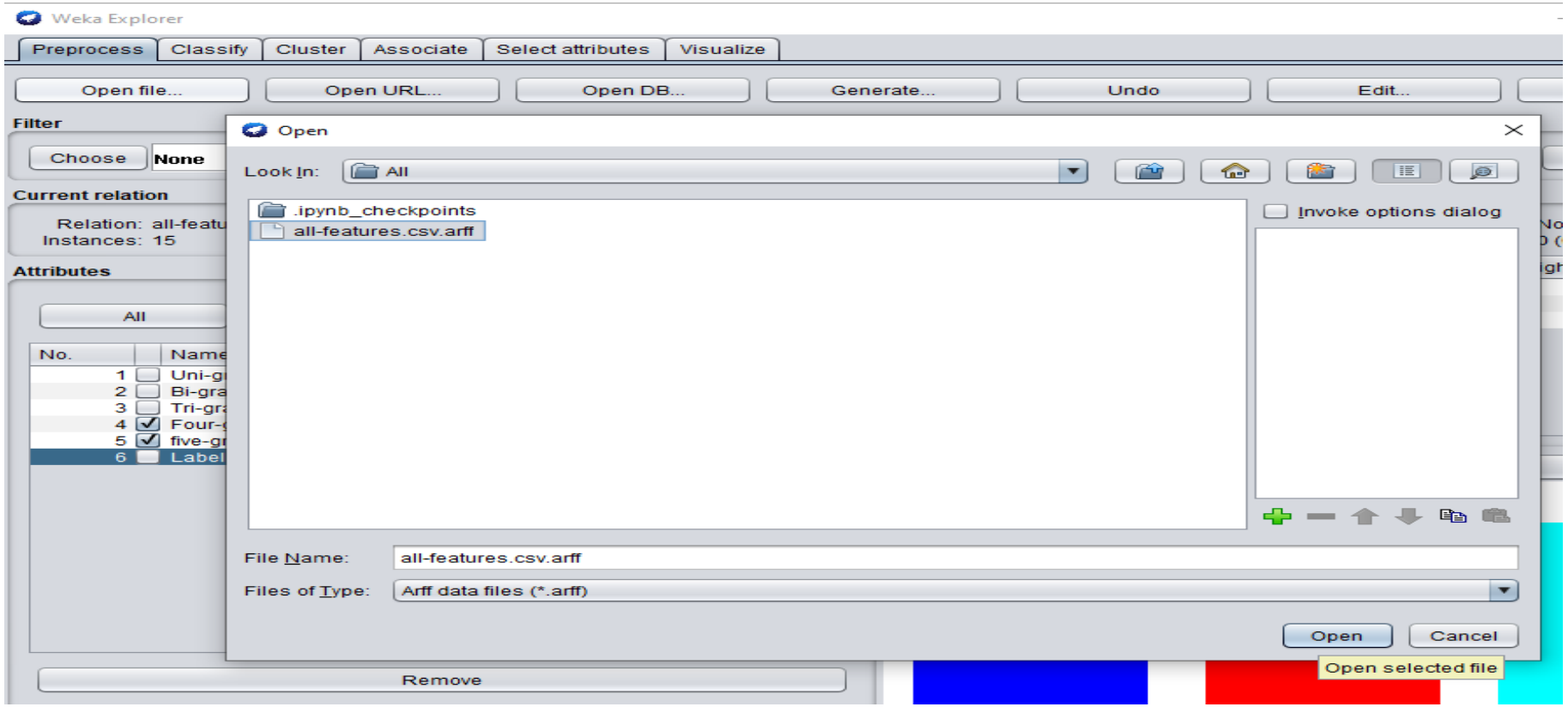
- Load features.csv
  - Click open



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**

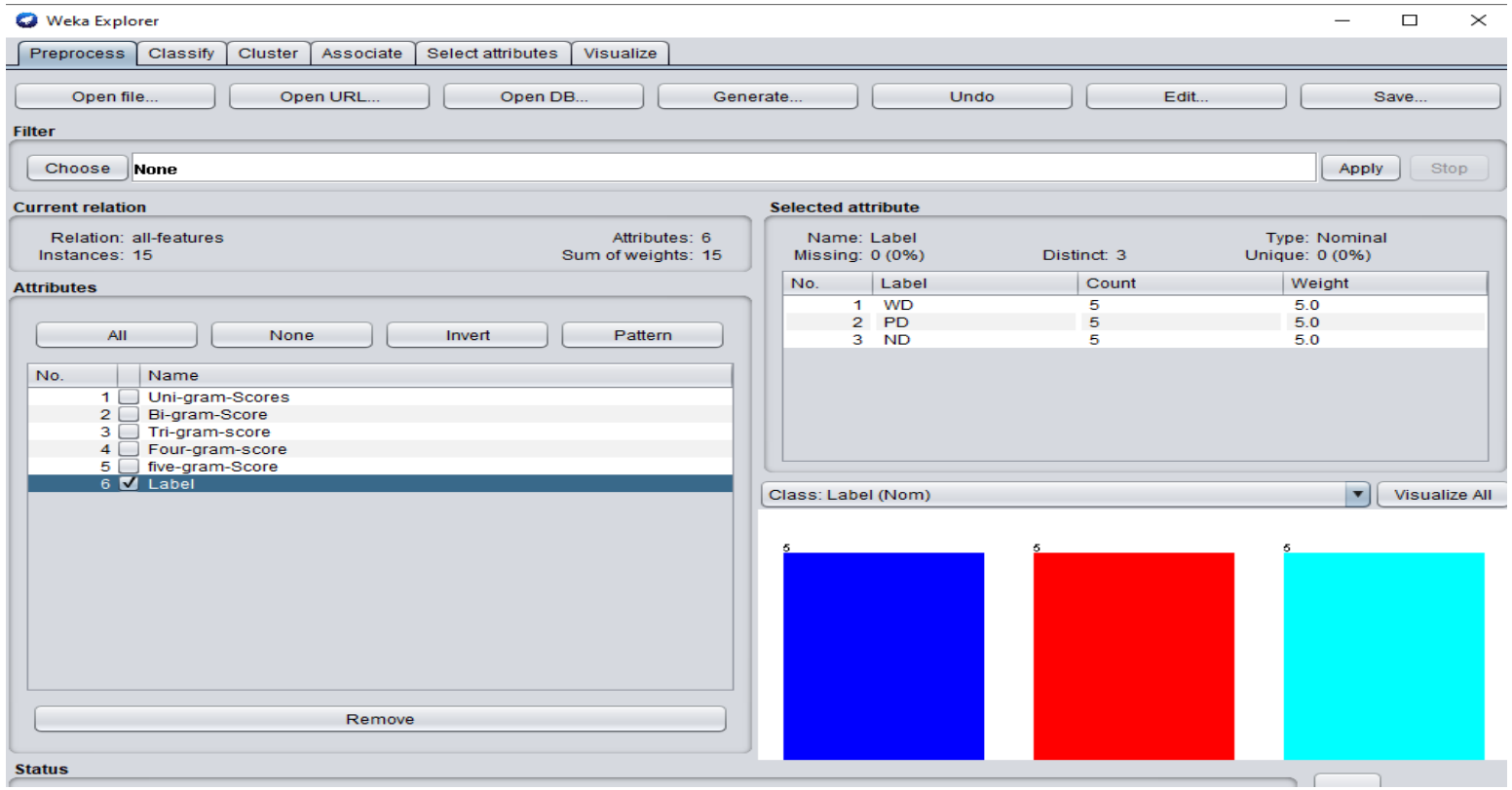


Example - Treating the Problem of Text Reuse Detection as Machine Learning Problem

Ternary Classification using WEKA

Load features.csv

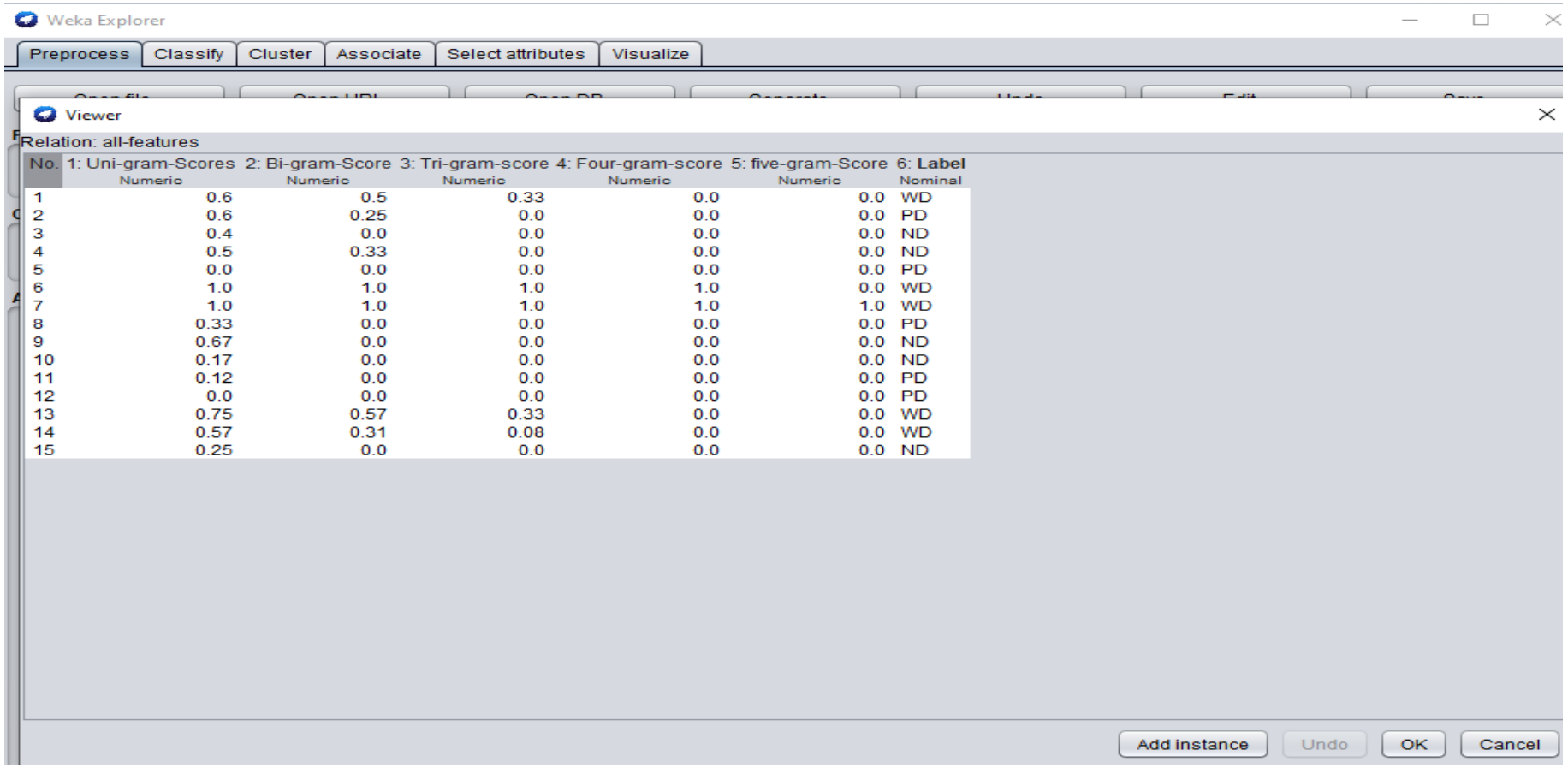
- Click Label and see Number of Classes



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

Ternary Classification using WEKA

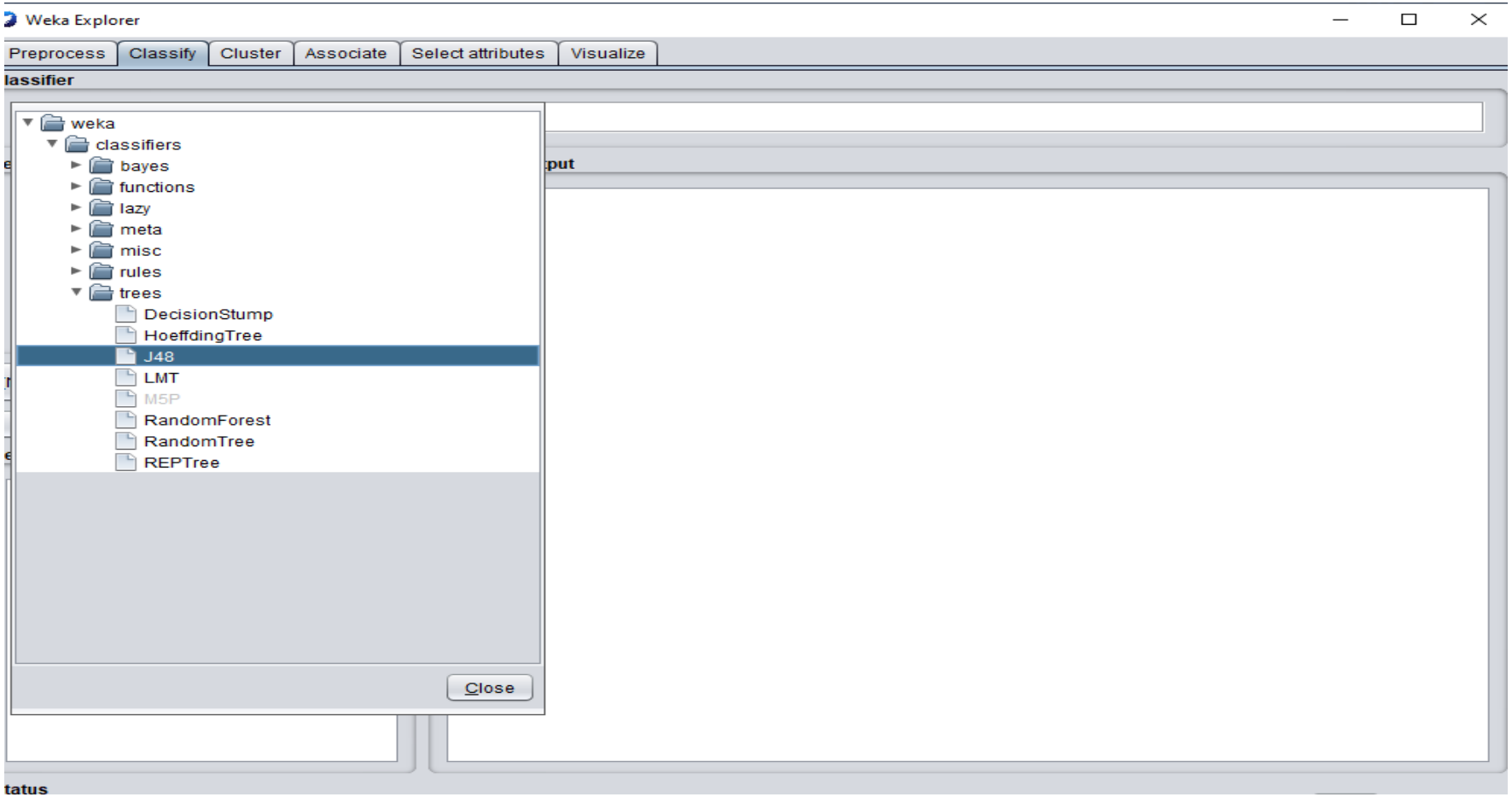
- Load features.csv
  - Click Edit and see Data View in Weka



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

Ternary Classification using WEKA

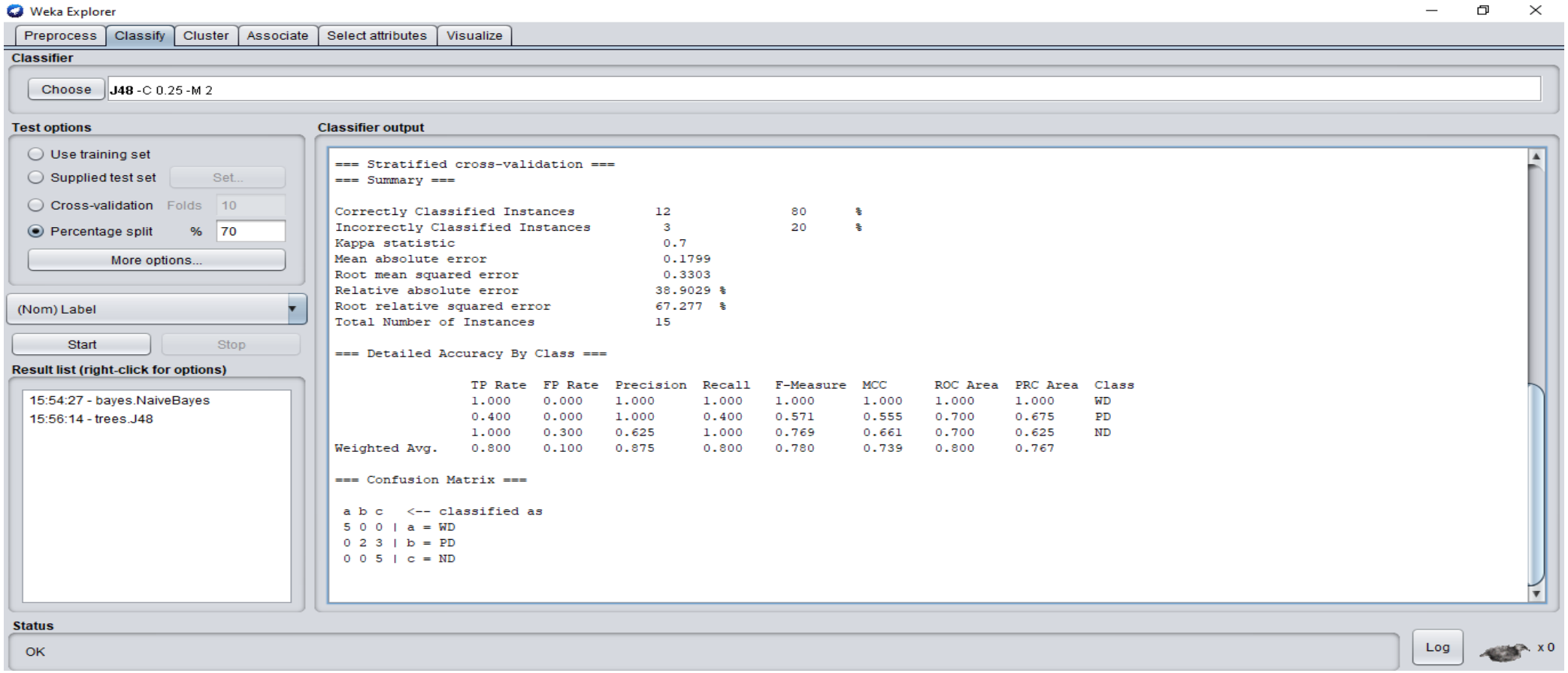
- Select J48



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

Ternary Classification using WEKA

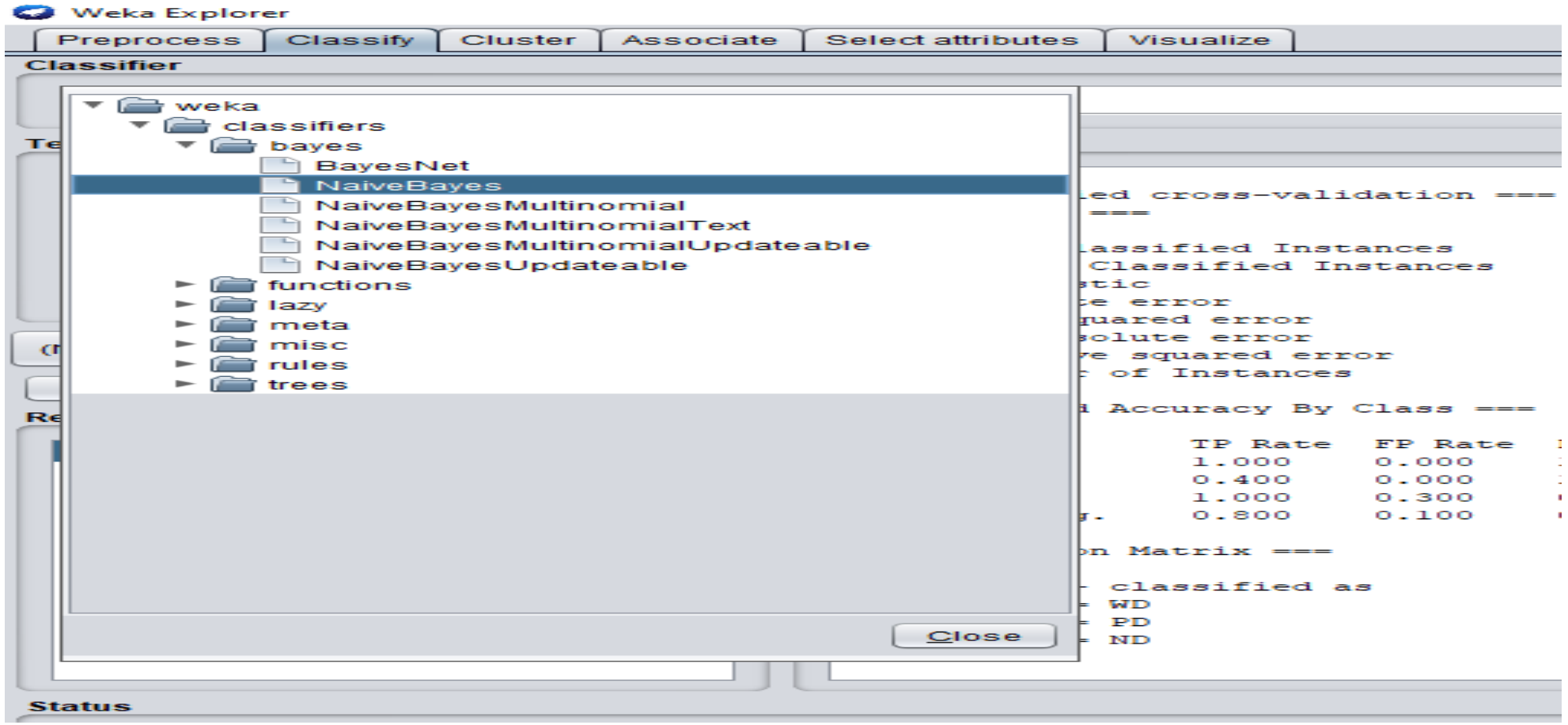
- Run J48 with Spilt ratio 70



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

Ternary Classification using WEKA

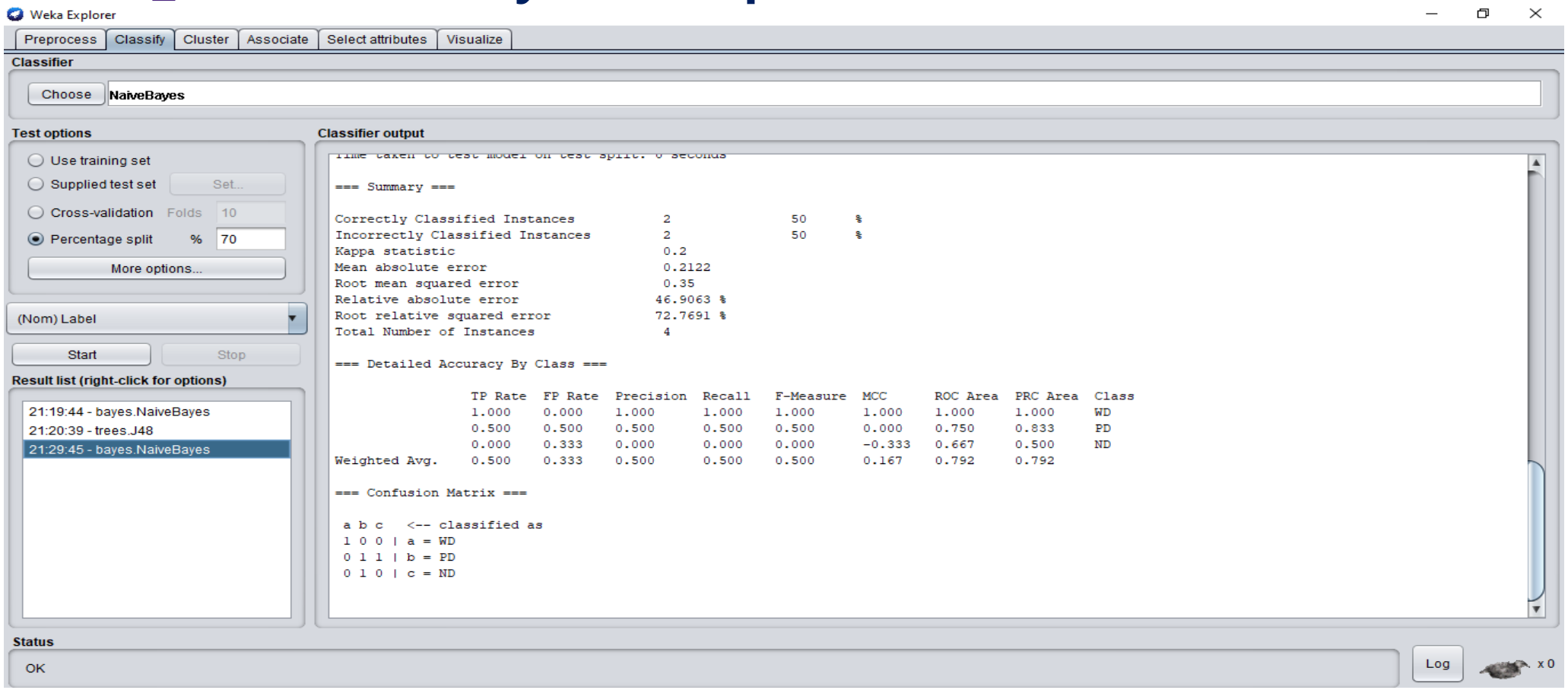
- Select Naive Bayes



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

Ternary Classification using WEKA

Run Naive Bayes with Spilt Ratio 70



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

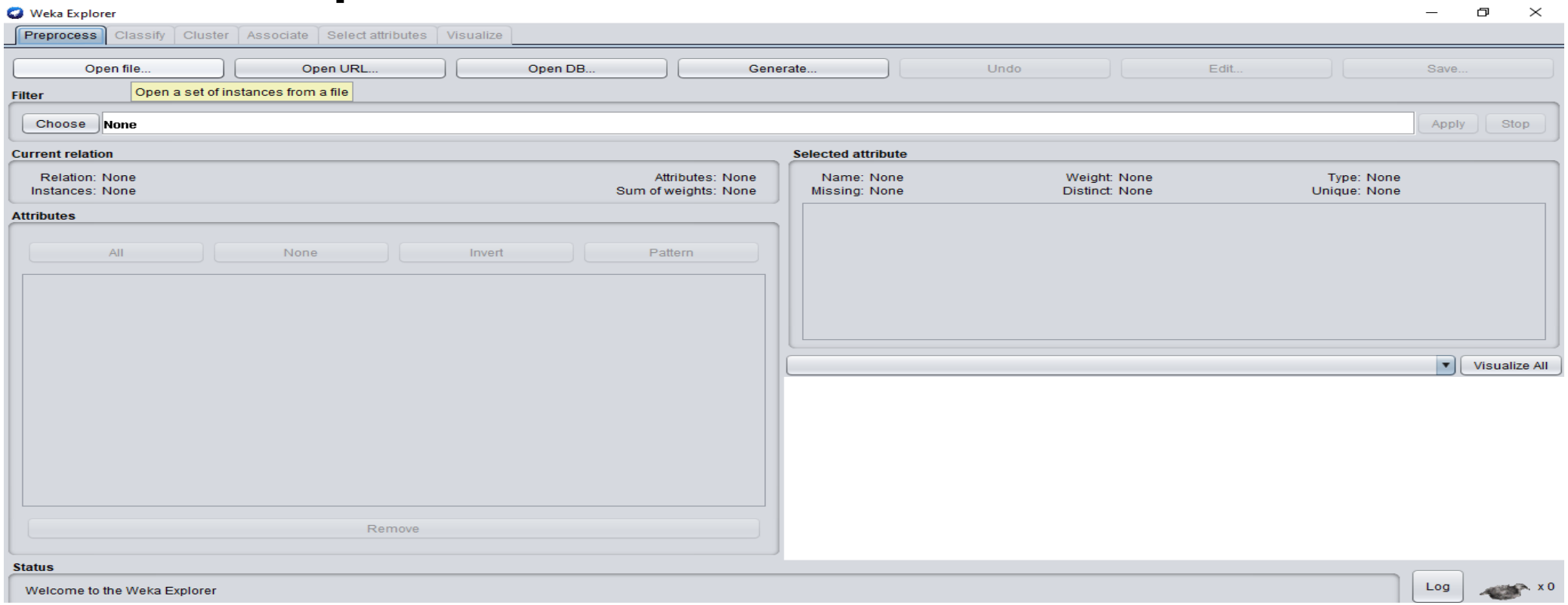
Your Turn

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.

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

Binary Classification using WEKA

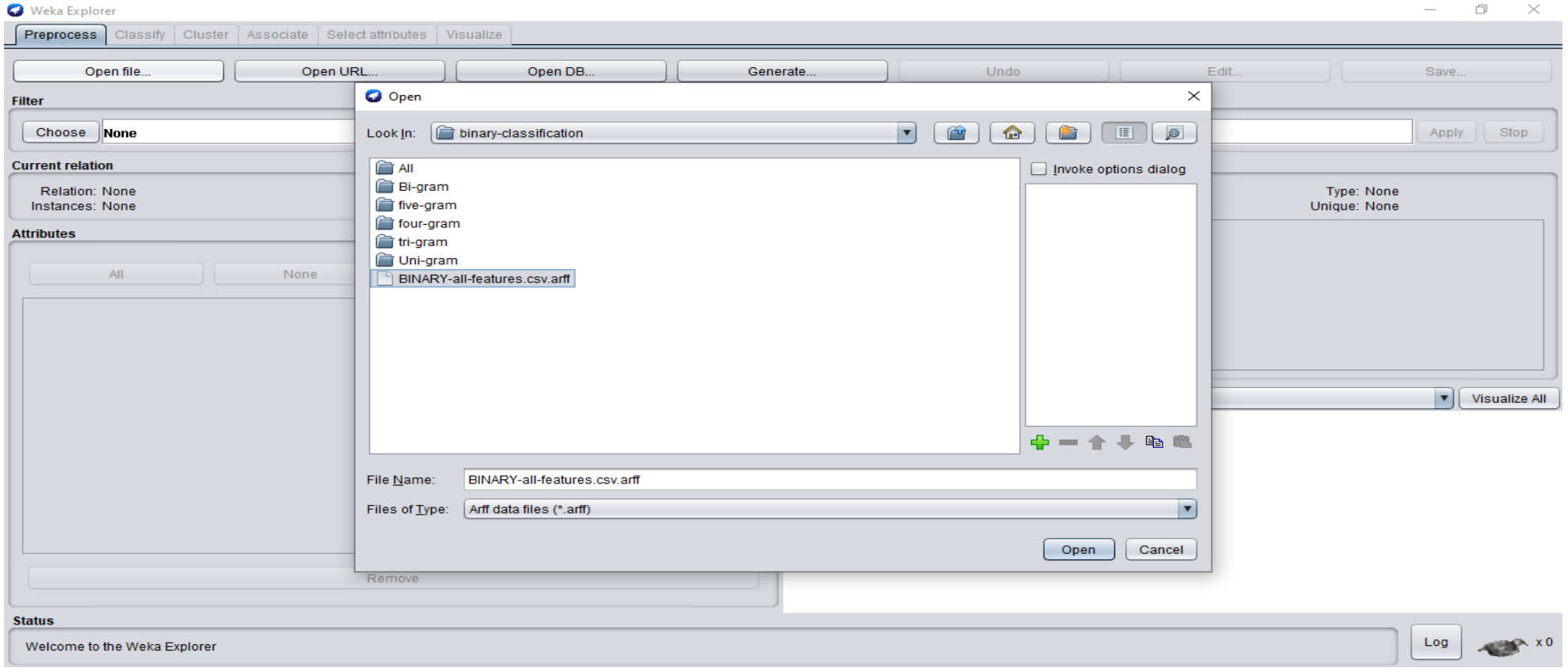
- Load features.csv
  - » Click Open



Example – (Cont.)

Binary Classification using WEKA

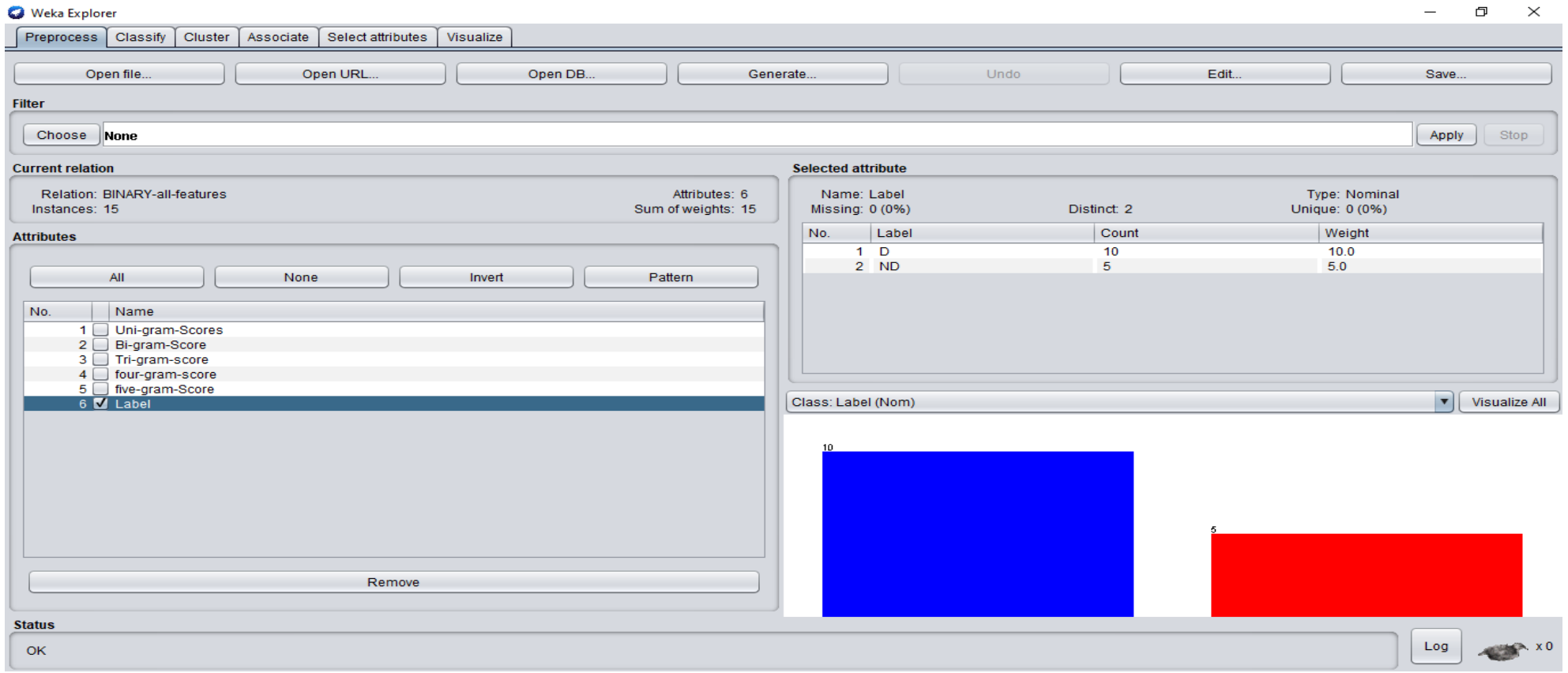
- Load features.csv
  - Select file by Giving Path



Example – (Cont.)

Binary Classification using WEKA

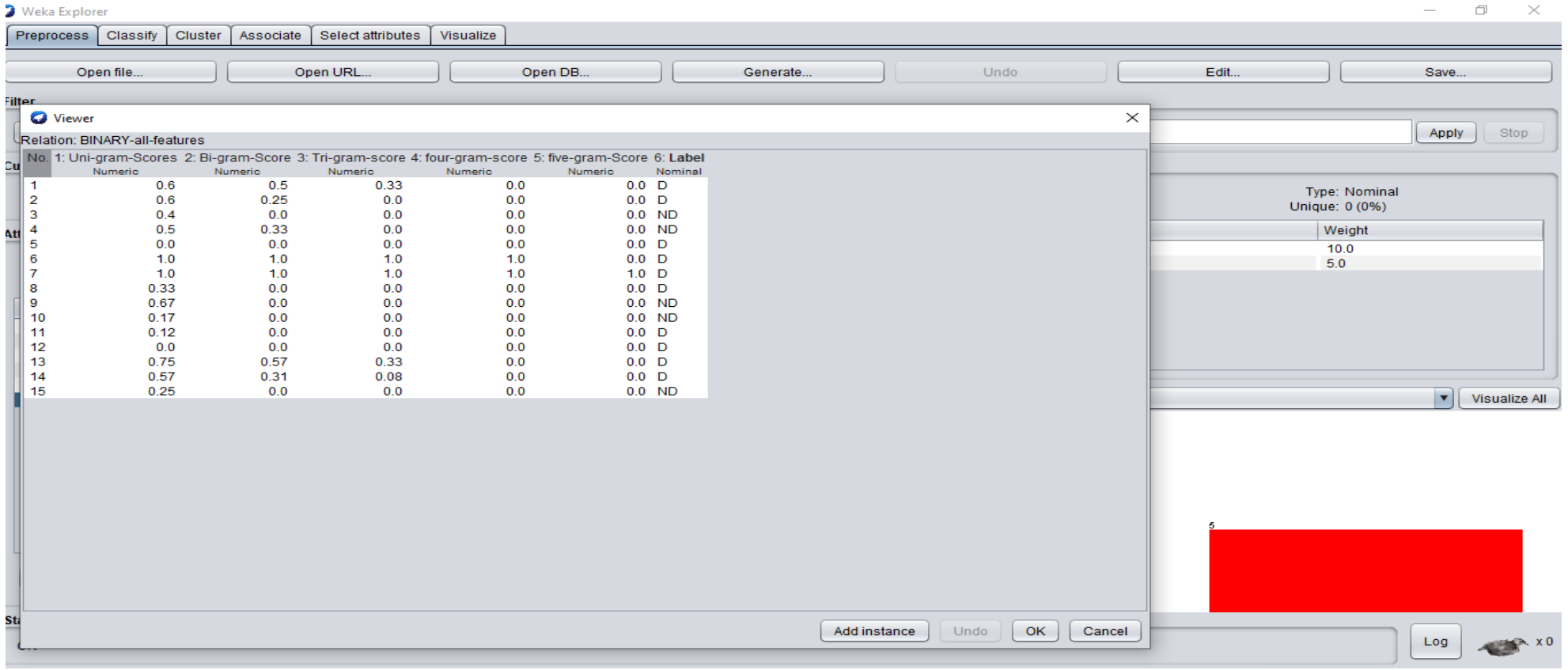
- Load features.csv
  - Click Label and see Number of Classes



Example – (Cont.)

Binary Classification using WEKA

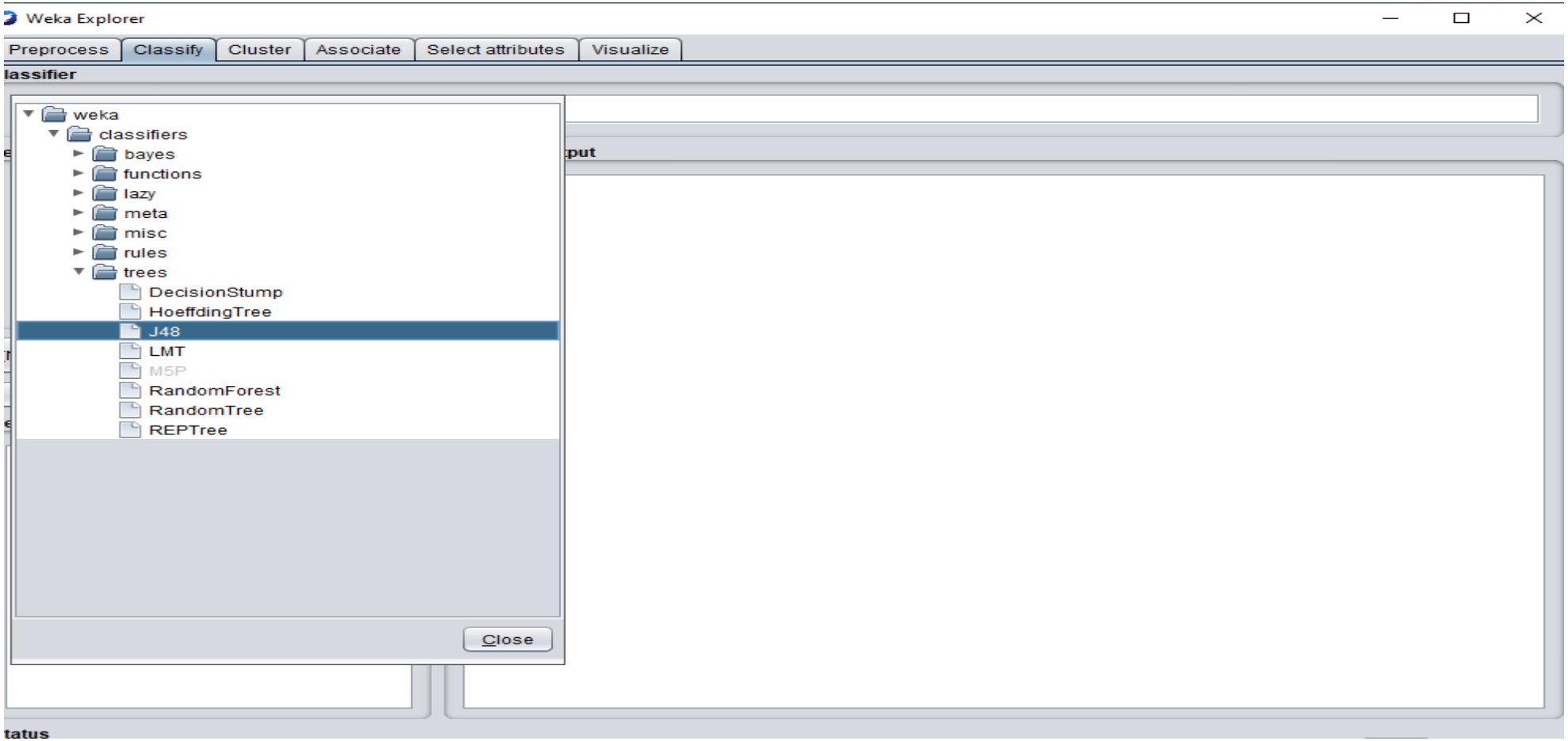
- Load features.csv
  - Click Edit and see Data View in WEKA



Example – (Cont.)

Binary Classification using WEKA

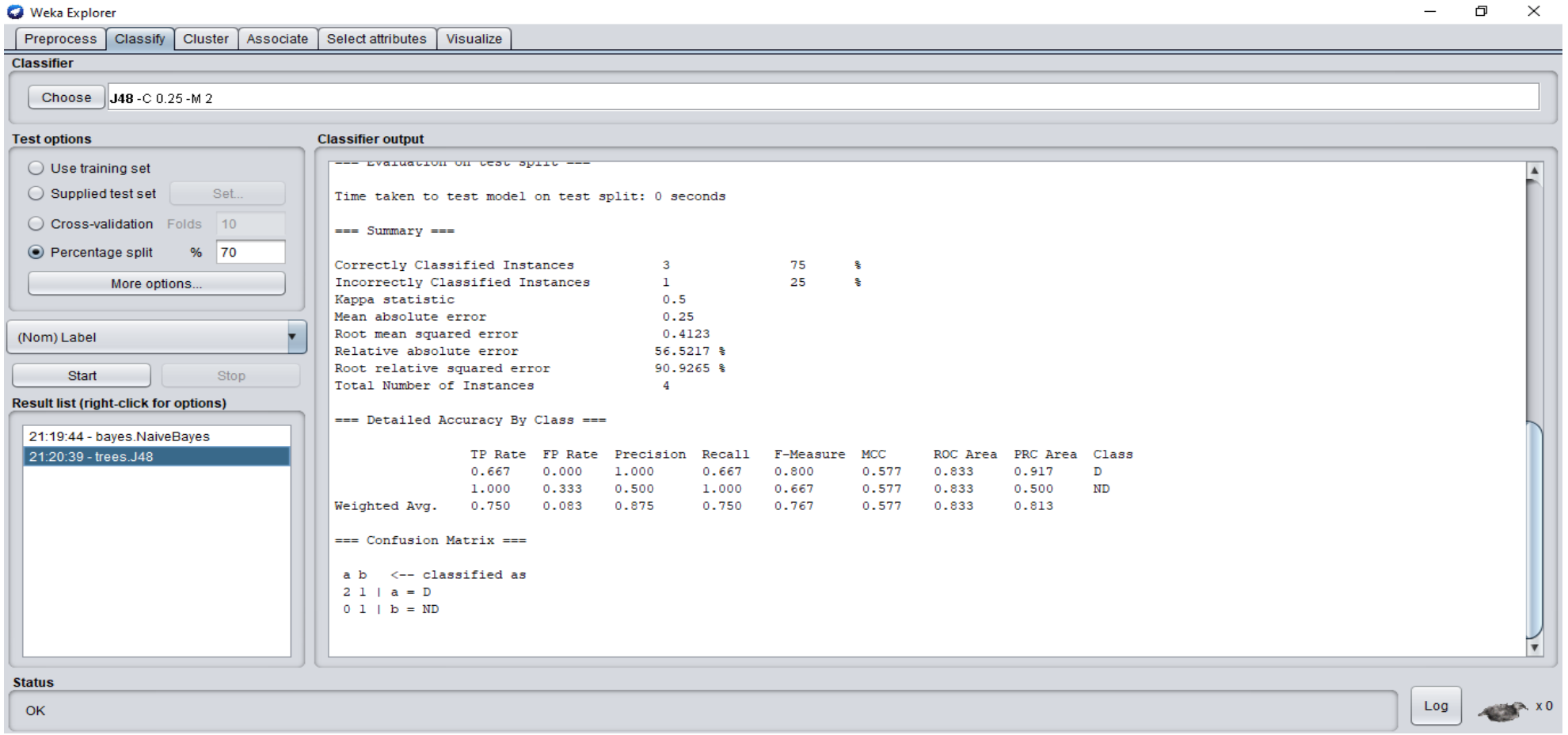
- Select J48



Example – (Cont.)

Binary Classification using WEKA

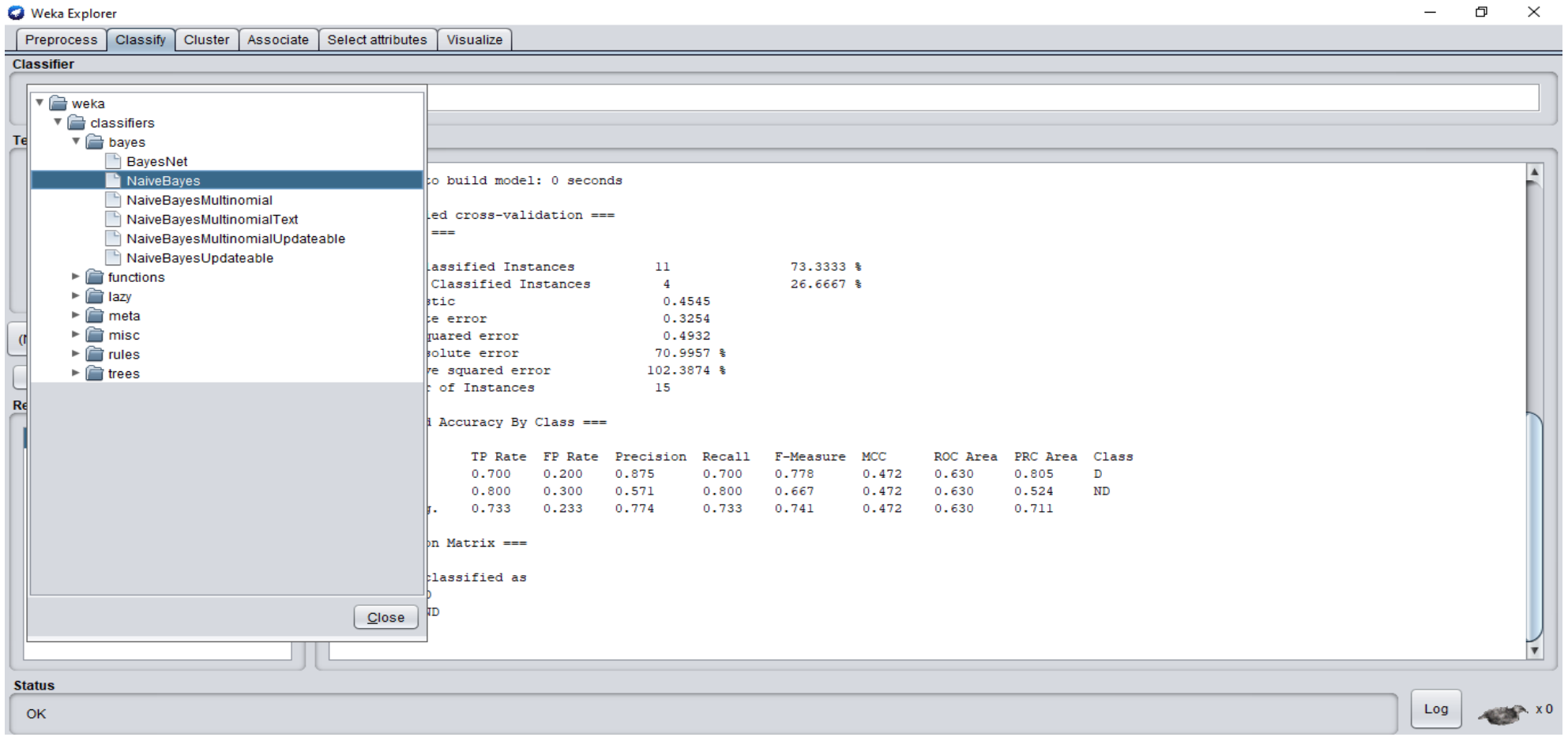
- Run J48 with Spilt ratio 70



Example – (Cont.)

Binary Classification using WEKA

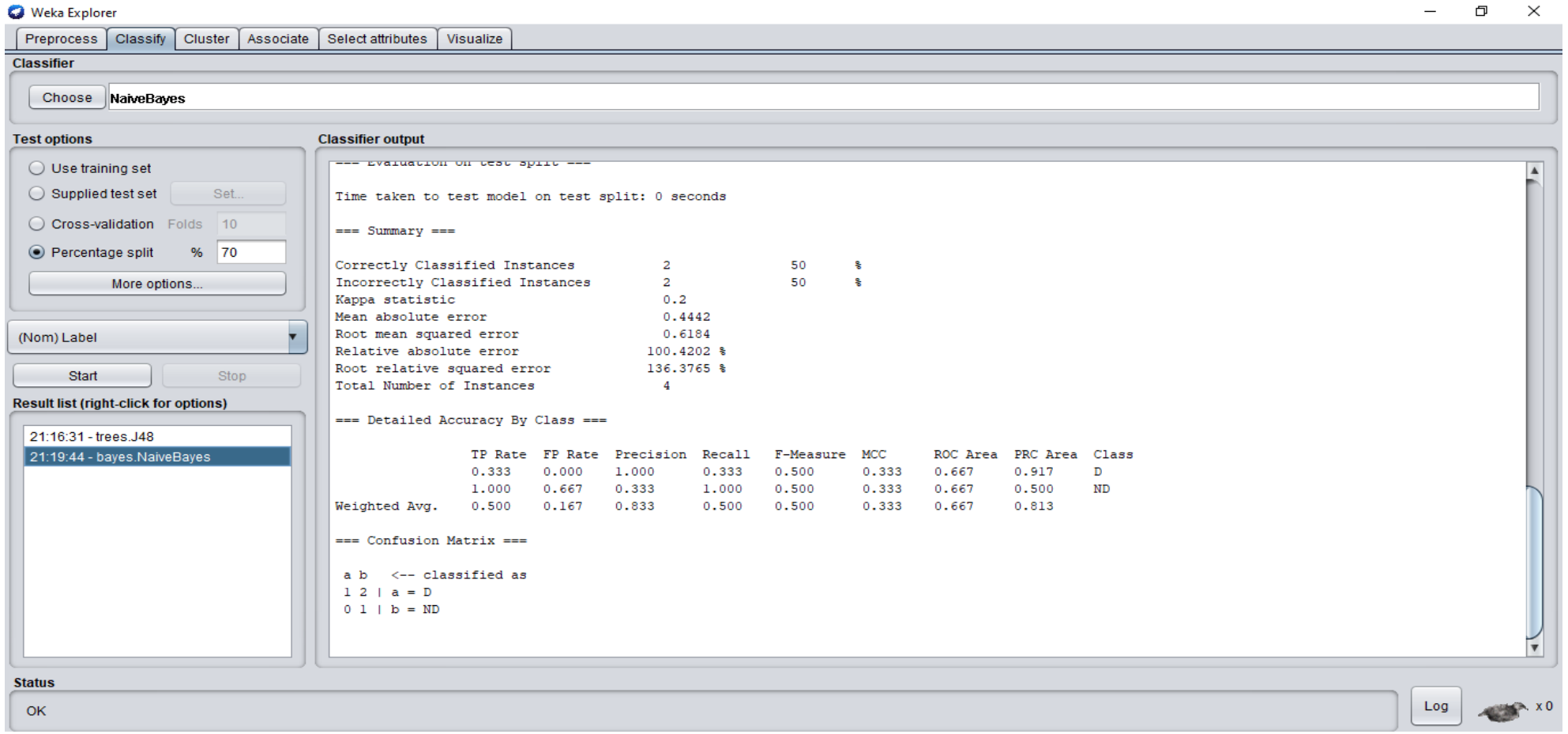
- Select Naive Bayes



Example – (Cont.)

Binary Classification using WEKA

- Run Naive Bayes with Spilt Ratio 70



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

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

Techniques(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

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

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)

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

Machine Learning Algorithms

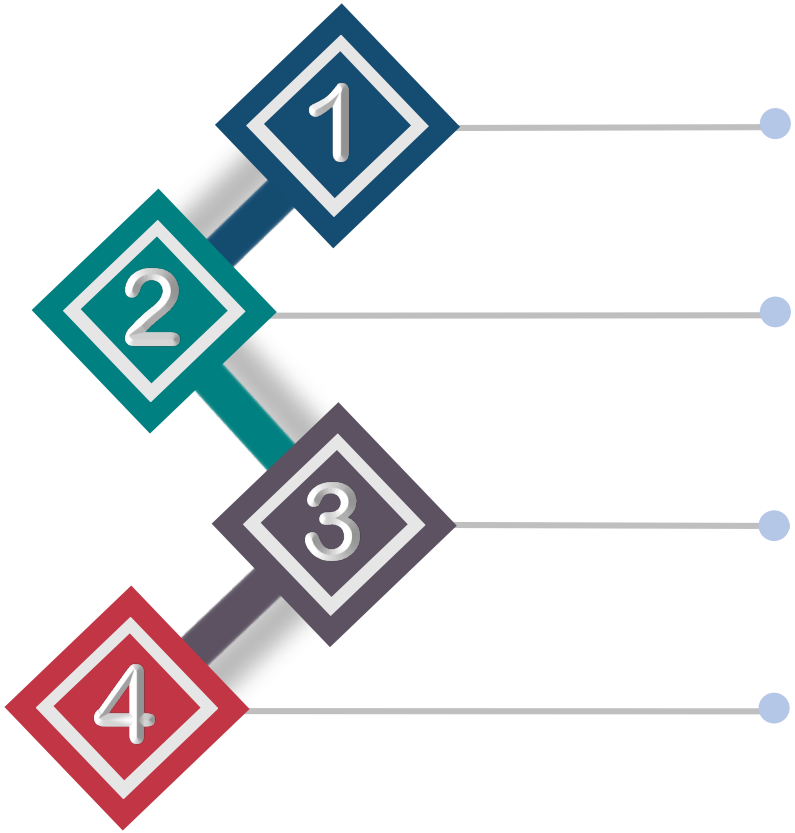
- Machine Learning Algorithms that will be trained / tested on features (similarity scores) extracted from the dataset
- For Supervised Text Classification task some of the Machine Learning Algorithms which have proven to be effective are

ML Algorithms

- Naïve Bayes
- Logistic Regression
- Multi-Layer Perceptron
- Random Forest
- Support Vector Machine
- AdaBoost

Summary: Treating the Problem of Text Reuse / Plagiarism Detection as Machine Learning Problem – A Step by Step Example (Cont.)

Main Steps to Treat Text Reuse Detection as Supervised Text Classification Task



1. **Pre-process input (i.e. text pair)**
2. **Feature Extraction from Input**
3. **Represent extracted features and output / label into a format that Machine Learning algorithms can understand (normally saved as CSV file)**
4. **Use features extracted in Step 2 to train and test Machine Learning algorithms**

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**

Summary – Introduction to Text Reuse and Plagiarism

- 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
- 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

Summary – Introduction to Text Reuse and Plagiarism

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

Summary – Introduction to Text Reuse and Plagiarism

- 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,

Summary – Introduction to Text Reuse and 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 Plagiarism** - 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

Summary – Introduction to Text Reuse and Plagiarism

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

Summary – Introduction to Text Reuse and Plagiarism

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

Summary – Introduction to Text Reuse and Plagiarism

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 Embedding's and (7) Methods based on Deep Learning**

Summary – Introduction to Text Reuse and Plagiarism

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

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

Techniques(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

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

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)

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

Machine Learning Algorithms

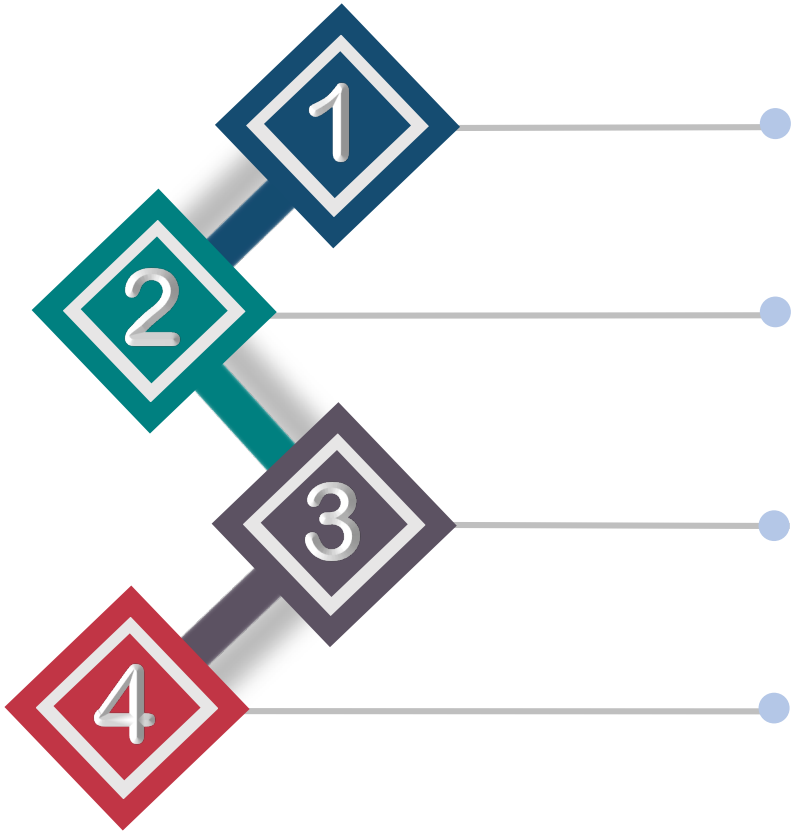
- Machine Learning Algorithms that will be trained / tested on features (similarity scores) extracted from the dataset
- For Supervised Text Classification task some of the Machine Learning Algorithms which have proven to be effective are

ML Algorithms

- Naïve Bayes
- Logistic Regression
- Multi-Layer Perceptron
- Random Forest
- Support Vector Machine
- AdaBoost

Summary: Treating the Problem of Text Reuse / Plagiarism Detection as Machine Learning Problem – A Step by Step Example (Cont.)

Main Steps to Treat Text Reuse Detection as Supervised Text Classification Task



1. **Pre-process input (i.e. text pair)**
2. **Feature Extraction from Input**
3. **Represent extracted features and output / label into a format that Machine Learning algorithms can understand (normally saved as CSV file)**
4. **Use features extracted in Step 2 to train and test Machine Learning algorithms**