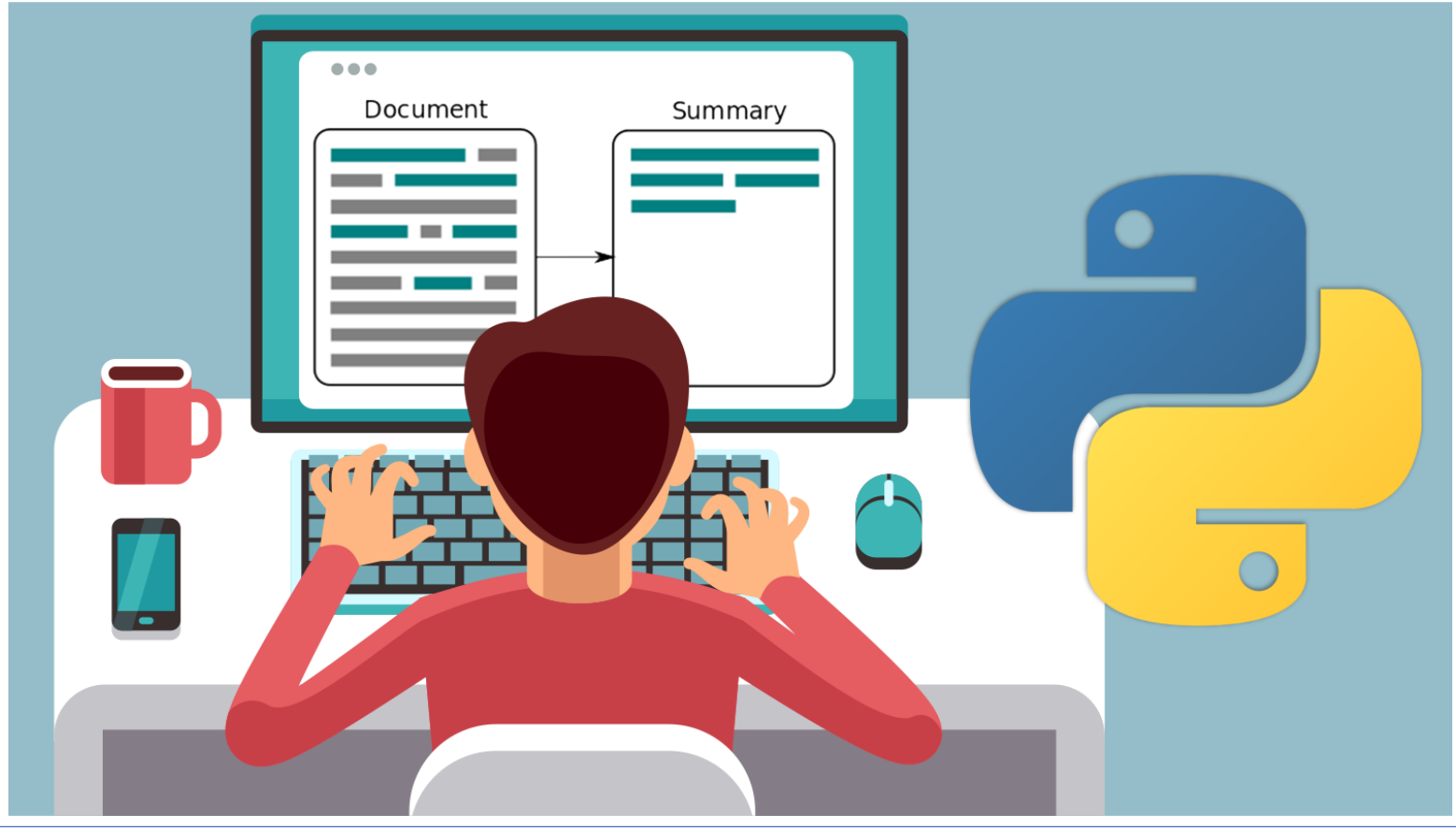
NATURAL LANGUAGE PROCESSING FOR TEXT SUMMARIZATION



EXPERT ACADEMY

COURSE CONTENT

- Frequency-based algorithm
- Luhn algorithm
- Cosine similarity
- Libraries
  - sumy
  - pysummarization
  - BERT summarizer
- **NLP** x Deep Learning
- Abstractive x **Extractive**

Summary - Automatic summarization - Wikipedia

Automatic summarization is the process of shortening a set of data computationally, to create a subset (a summary) that represents the most important or relevant information within the original content. In addition to text, images and videos can also be summarized. Text summarization finds the most informative sentences in a document;[1] image summarization finds the most representative images within an image collection[citation needed]; video summarization extracts the most important frames from the video content. [2] There are two general approaches to automatic summarization: extraction and abstraction. Here, content is extracted from the original data, but the extracted content is not modified in any way. Examples of extracted content include key-phrases that can be used to "tag" or index a text document, or key sentences (including headings) that collectively comprise an abstract, and representative images or video segments, as stated above. For text, extraction is analogous to the process of skimming, where the summary (if available), headings and subheadings, figures, the first and last paragraphs of a section, and optionally the first and last sentences in a paragraph are read before one chooses to read the entire document in detail. [3] Other examples of extraction that include key sequences of text in terms of clinical relevance (including patient/problem, intervention, and outcome). [4] This has been applied mainly for text. Abstractive methods build an internal semantic representation of the original content, and then use this representation to create a summary that is closer to what a human might express. Abstraction may transform the extracted content by paraphrasing sections of the source document, to condense a text more strongly than extraction. Such transformation, however, is computationally much more challenging than extraction, involving both natural language processing and often a deep understanding of the domain of the original text in cases where the original document relates to a special field of knowledge. "Paraphrasing" is even more difficult to apply to image and video, which is why most summarization systems are extractive. Approaches aimed at higher summarization quality rely on combined software and human effort. In Machine Aided Human Summarization, extractive techniques highlight candidate passages for inclusion (to which the human adds or removes text). In Human Aided Machine Summarization, a human post-processes software output, in the same way that one edits the

PREREQUISITES

- Programming logic
- Basic Python programming
- Level: **begginners**

NATURAL LANGUAGE PROCESSING

Speech Transcription

Neural Machine Translation (NMT)

Chatbots

Q&A

Text Summarization

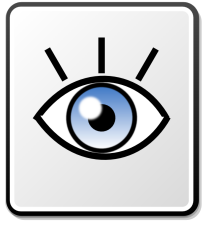


Image Captioning

Video Captioning

Sentiment analysis

PLAN OF ATTACK – FREQUENCY BASED ALGORITHM

1. Intuition and step by step calculations
2. Step by step implementation in Python
3. Generate sumaries using articles from the Internet

Summary - Lemmatisation - Wikipedia

Lemmatisation (or lemmatization) in linguistics is the process of grouping together the inflected forms of a word so they can be analysed as a single item, identified by the word's lemma, or dictionary form. [1] In computational linguistics, lemmatisation is the algorithmic process of determining the lemma of a word based on its intended meaning. Unlike stemming, lemmatisation depends on correctly identifying the intended part of speech and meaning of a word in a sentence, as well as within the larger context surrounding that sentence, such as neighboring sentences or even an entire document. As a result, developing efficient lemmatisation algorithms is an open area of research. [2][3][4] In many languages, words appear in several inflected forms. For example, in English, the verb 'to walk' may appear as 'walk', 'walked', 'walks' or 'walking'. The base form, 'walk', that one might look up in a dictionary, is called the lemma for the word. The association of the base form with a part of speech is often called a lexeme of the word. Lemmatisation is closely related to stemming. The difference is that a stemmer operates on a single word without knowledge of the context, and therefore cannot discriminate between words which have different meanings depending on part of speech. However, stemmers are typically easier to implement and run faster. The reduced "accuracy" may not matter for some applications. In fact, when used within information retrieval systems, stemming improves query recall accuracy, or true positive rate, when compared to lemmatisation. Nonetheless, stemming reduces precision, or the proportion of positively-labeled instances that are actually positive, for such systems. [5] • The word "better" has "good" as its lemma. This link is missed by stemming, as it requires a dictionary look-up. • The word "walk" is the base form for the word "walking", and hence this is matched in both stemming and lemmatisation. • The word "meeting" can be either the base form of a noun or a form of a verb ("to meet") depending on the context; e.g., "in our last meeting" or "We are meeting again tomorrow". Unlike stemming, lemmatisation attempts to select the correct lemma depending on the context. Document indexing software like Lucene[6] can store the base stemmed format of the word without the knowledge of meaning, but only considering word formation grammar rules. The stemmed word itself might not be a valid word: 'lazy', as seen in the example below, is stemmed by many stemmers to 'lazi'. This is because the purpose of stemming is not to produce the appropriate lemma – that is a more challenging task that requires knowledge of context. The main purpose of stemming is to map different forms of a word to a single form. [7] As a rule-based algorithm, dependent only upon the spelling of a word, it sacrifices accuracy to ensure that, for example, when 'laziness' is stemmed to 'lazi', it has the same stem as 'lazy'. A trivial way to do lemmatization is by simple dictionary lookup. This works well for straightforward inflected forms, but a rule-based system will be needed for other cases, such as in languages with long compound words. Such rules can be either hand-crafted or learned automatically from an annotated corpus. Morphological analysis of published biomedical literature can yield useful results. Morphological processing of biomedical text can be more effective by a specialised lemmatisation program for biomedicine, and may improve the accuracy of practical information extraction tasks. [8]

TEXT SUMMARIZATION

- Artificial intelligence is human like intelligence. It is the study of intelligent artificial agents. Science and engineering to produce intelligent machines. Solve problems and have intelligence. Related to intelligent behavior. Developing of reasoning machines. Learn from mistakes and successes. Artificial intelligence is related to reasoning in everyday situations.
- Steps
    1. Preprocessing the texts
    2. Word frequency
    3. Weighted word frequency
    4. Sentence tokenization
    5. Score for the sentences
    6. Order the sentences
    7. Generate the summary

1. PREPROCESSING THE TEXTS

- Artificial intelligence is human like intelligence. It is the study of intelligent artificial agents. Science and engineering to produce intelligent machines. Solve problems and have intelligence. Related to intelligent behavior. Developing of reasoning machines. Learn from mistakes and successes. Artificial intelligence is related to reasoning in everyday situations

- artificial intelligence human like intelligence. study intelligent artificial agents. science engineering produce intelligent machines. solve problems intelligence. related intelligent behavior. developing reasoning machines. learn mistakes successes. artificial intelligence related reasoning everyday situations

2. WORD FREQUENCY

Word	Frequency
artificial	3
intelligence	4
human	1
like	1
study	1
intelligent	3
science	1
engineering	1
produce	1
machines	2
solve	1

Word	Frequency
agents	1
problems	1
related	2
behavior	1
developing	1
reasoning	2
learn	1
mistakes	1
successes	1
everyday	1
situations	1

3. WEIGHTED WORD FREQUENCY

Highest value: 4

Word	Frequency	Weigth
artificial	3	0.75
intelligence	4	1.00
human	1	0.25
like	1	0.25
study	1	0.25
intelligent	3	0.75
science	1	0.25
engineering	1	0.25
produce	1	0.25
machines	2	0.50
solve	1	0.25

Word	Frequency	Weight
agents	1	0.25
problems	1	0.25
related	2	0.50
behavior	1	0.25
developing	1	0.25
reasoning	2	0.50
learn	1	0.25
mistakes	1	0.25
successes	1	0.25
everyday	1	0.25
situations	1	0.25

4. SENTENCE TOKENIZATION

- Artificial intelligence is human like intelligence. It is the study of intelligent artificial agents. Science and engineering to produce intelligent machines. Solve problems and have intelligence. Related to intelligent behavior. Developing of reasoning machines. Learn from mistakes and successes. Artificial intelligence is related to reasoning in everyday situations
- Tokenization
  - Artificial intelligence is human like intelligence.
  - It is the study of intelligent artificial agents.
  - Science and engineering to produce intelligent machines.
  - Solve problems and have intelligence.
  - Related to intelligent behavior.
  - Developing of reasoning machines.
  - Learn from mistakes and successes.
  - Artificial intelligence is related to reasoning in everyday situations

5. SCORE FOR THE SENTENCES

Sentence	Score (sum of weights)
Artificial **(0.75)** intelligence **(1.00)** is human **(0.25)** like **(0.25)** intelligence **(1.00)**.	3.25
It is the study **(0.25)** of intelligent **(0.75)** artificial **(0.75)** agents **(0.25)**.	2.00
Science **(0.25)** and engineering **(0.25)** to produce **(0.25)** intelligent **(0.75)** machines **(0.50)**.	2.00
Solve **(0.25)** problems **(0.25)** and have intelligence **(1.00)**.	1.50
Related **(0.50)** to intelligent **(0.75)** behavior **(0.25)**.	1.50
Developing **(0.25)** of reasoning **(0.50)** machines **(0.50)**.	1.25
Learn **(0.25)** from mistakes **(0.25)** and successes **(0.25)**.	0.75
Artificial **(0.75)** intelligence **(1.00)** is related **(0.50)** to reasoning **(0.50)** in everyday **(0.25)** situations **(0.25)**.	3.25

6. ORDER THE SENTENCES

Sentence	Score (sum of weights)
Artificial **(0.75)** intelligence **(1.00)** is related **(0.50)** to reasoning **(0.50)** in everyday **(0.25)** situations **(0.25)**.	3.25
Artificial **(0.75)** intelligence **(1.00)** is human **(0.25)** like **(0.25)** intelligence **(1.00)**.	3.25
It is the study **(0.25)** of intelligent **(0.75)** artificial **(0.75)** agents **(0.25)**.	2.00
Science **(0.25)** and engineering **(0.25)** to produce **(0.25)** intelligent **(0.75)** machines **(0.50)**.	2.00
Solve **(0.25)** problems **(0.25)** and have intelligence **(1.00)**.	1.50
Related **(0.50)** to intelligent **(0.75)** behavior **(0.25)**.	1.50
Developing **(0.25)** of reasoning **(0.50)** machines **(0.50)**.	1.25
Learn **(0.25)** from mistakes **(0.25)** and successes **(0.25)**.	0.75

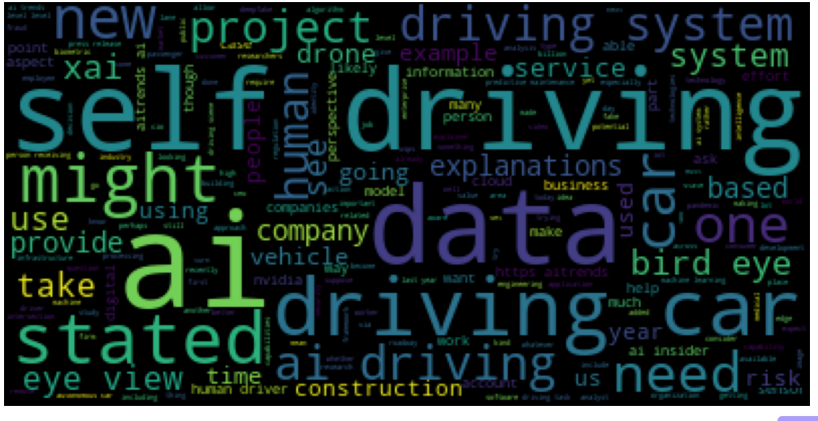
7. GENERATE THE SUMMARY

- Artificial intelligence is human like intelligence. It is the study of intelligent artificial agents. Science and engineering to produce intelligent machines. Solve problems and have intelligence. Related to intelligent behavior. Developing of reasoning machines. Learn from mistakes and successes. Artificial intelligence is related to reasoning in everyday situations

- Artificial intelligence is related to reasoning in everyday situations. Artificial intelligence is human like intelligence. It is the study of intelligent artificial agents.

LUHN ALGORITHM

1. Basic intuition
2. Step by step implementation in Python
3. Reading from news feed



Summary - Automatic summarization - Wikipedia

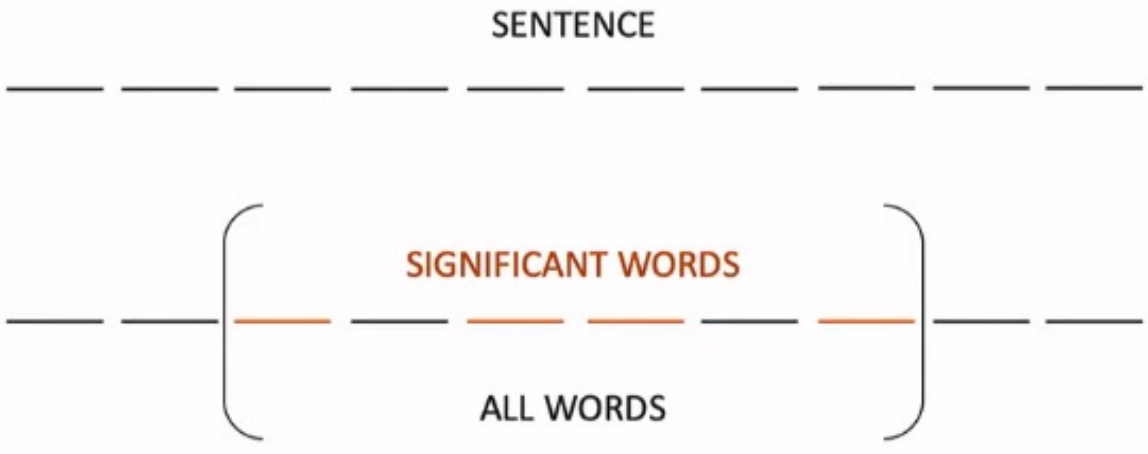
Automatic summarization is the process of shortening a set of data computationally, to create a subset (a summary) that represents the most important or relevant information within the original content. In addition to text, images and videos can also be summarized. Text summarization finds the most informative sentences in a document;[1] image summarization finds the most representative images within an image collection[citation needed]; video summarization extracts the most important frames from the video content. [2] There are two general approaches to automatic summarization: extraction and abstraction. Here, content is extracted from the original data, but the extracted content is not modified in any way. Examples of extracted content include key-phrases that can be used to "tag" or index a text document, or key sentences (including headings) that collectively comprise an abstract, and representative images or video segments, as stated above. For text, extraction is analogous to the process of skimming, where the summary (if available), headings and subheadings, figures, the first and last paragraphs of a section, and optionally the first and last sentences in a paragraph are read before one chooses to read the entire document in detail. [3] Other examples of extraction that include key sequences of text in terms of clinical relevance (including patient/problem, intervention, and outcome). [4] This has been applied mainly for text. Abstractive methods build an internal semantic representation of the original content, and then use this representation to create a summary that is closer to what a human might express. Abstraction may transform the extracted content by paraphrasing sections of the source document, to condense a text more strongly than extraction. Such transformation, however, is computationally much more challenging than extraction, involving both natural language processing and often a deep understanding of the domain of the original text in cases where the original document relates to a special field of knowledge. "Paraphrasing" is even more difficult to apply to image and video, which is why most summarization systems are extractive. Approaches aimed at higher summarization quality rely on combined software and human effort. In Machine Aided Human Summarization, extractive techniques highlight candidate passages for inclusion (to which the human adds or removes text). In Human Aided Machine Summarization, a human post-processes software output, in the same way that one edits the

john p. desmond PERSON ai trends editor predictive maintenance pdm emerged killer ai app past five years DATE predictive maintenance moved niche use case fast-growing high return investment roi application delivering true value users developments indication power internet things iot ai together market considered infancy today DATE observations research conducted iot analytics consultants supply market intelligence recently estimated 6.9 billion CARDINAL predictive maintenance market reach 28.2 billion MONEY 2026 DATE . company began research coverage iot-driven predictive maintenance market 2016 DATE industry maintenance conference dortmund germany GPE much happening “ bitterly disappointed ” stated knud lasse lueth PERSON ceo iot analytics account iot business news “ single exhibitor talking predictive maintenance. ” things changed iot analytics analyst fernando alberto brügge PERSON stated “ research 2021 shows predictive maintenance clearly

EXPERT ACADEMY

LUHN ALGORITHM

- Select most important words based on frequency
- Score calculation: $4^2/6 = 2,7$



Source: https://iq.opengenus.org/luhns-heuristic-method-for-text-summarization/

COSINE SIMILARITY

- Implementation in Python

Summary - Automatic summarization - Wikipedia

Automatic summarization is the process of shortening a set of data computationally, to create a subset (a summary) that represents the most important or relevant information within the original content. In addition to text, images and videos can also be summarized. Text summarization finds the most informative sentences in a document;[1] image summarization finds the most representative images within an image collection[citation needed]; video summarization extracts the most important frames from the video content. [2] There are two general approaches to automatic summarization: extraction and abstraction. Here, content is extracted from the original data, but the extracted content is not modified in any way. Examples of extracted content include key-phrases that can be used to "tag" or index a text document, or key sentences (including headings) that collectively comprise an abstract, and representative images or video segments, as stated above. For text, extraction is analogous to the process of skimming, where the summary (if available), headings and subheadings, figures, the first and last paragraphs of a section, and optionally the first and last sentences in a paragraph are read before one chooses to read the entire document in detail. [3] Other examples of extraction that include key sequences of text in terms of clinical relevance (including patient/problem, intervention, and outcome). [4] This has been applied mainly for text. Abstractive methods build an internal semantic representation of the original content, and then use this representation to create a summary that is closer to what a human might express. Abstraction may transform the extracted content by paraphrasing sections of the source document, to condense a text more strongly than extraction. Such transformation, however, is computationally much more challenging than extraction, involving both natural language processing and often a deep understanding of the domain of the original text in cases where the original document relates to a special field of knowledge. "Paraphrasing" is even more difficult to apply to image and video, which is why most summarization systems are extractive. Approaches aimed at higher summarization quality rely on combined software and human effort. In Machine Aided Human Summarization, extractive techniques highlight candidate passages for inclusion (to which the human adds or removes text). In Human Aided Machine Summarization, a human post-processes software output, in the same way that one edits the

TEXT SUMMARIZATION USING LIBRARIES

- sumy
- pysummarization
- BERT summarizer