NATURAL LANGUAGE PROCESSING

LESSON 12: KEYWORD EXTRACTION & DOCUMENT SUMMARIZATION

OUTLINE

- Keyword Extraction
- Keyword Extraction Approaches
  - ➢ Rule Based Linguistic Approaches
  - ➢ Machine Learning Approaches
  - ➢ Statistical Approaches
- Document Summarization
  - ➢ Extraction-based
  - ➢ Abstraction-based

THE PROBLEM

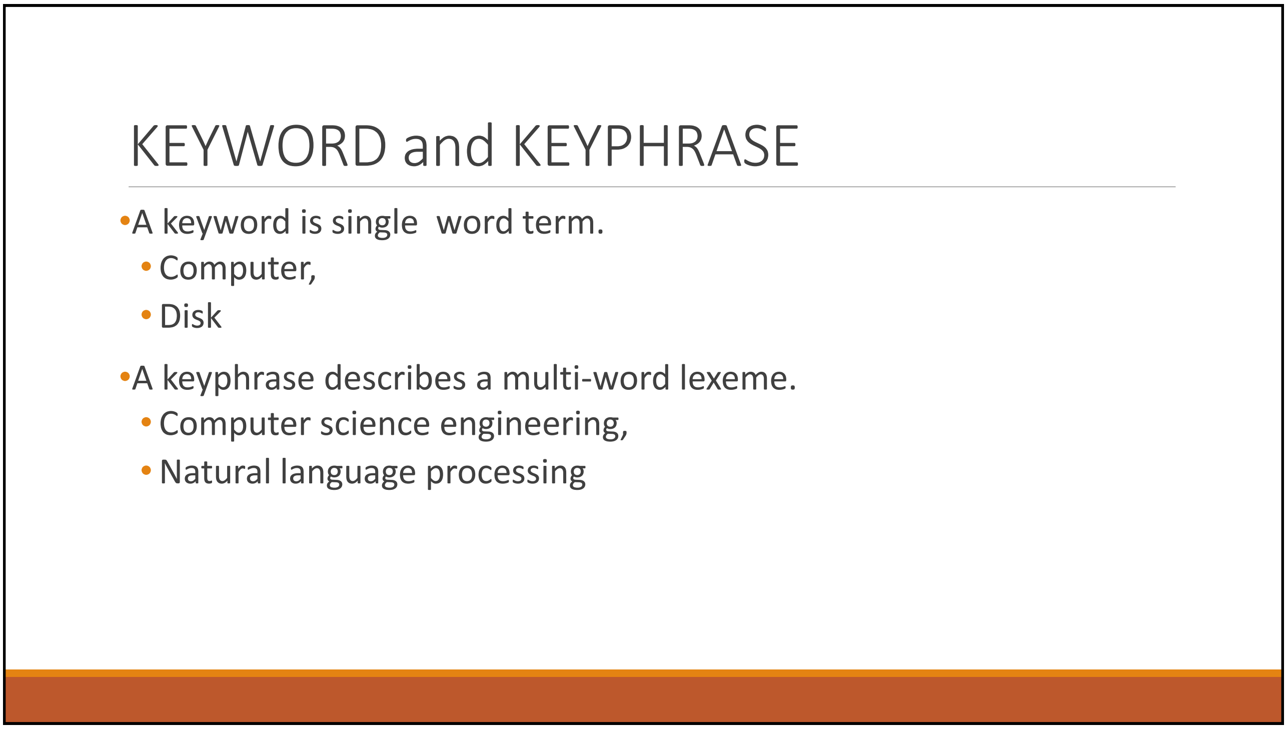
- Document **reading** is a **time consuming** task… 😐
- Many **common documents** (e.g., e-mail, newsgroup posts, web pages) **lack of abstract or keywords**…☹
- But, they are **“electronic”** so we can **work on them** in some way… ☺

KEYWORD EXTRACTION

Keyword extraction is defined as the task that automatically identifies a set of the terms that best describe the subject of document.

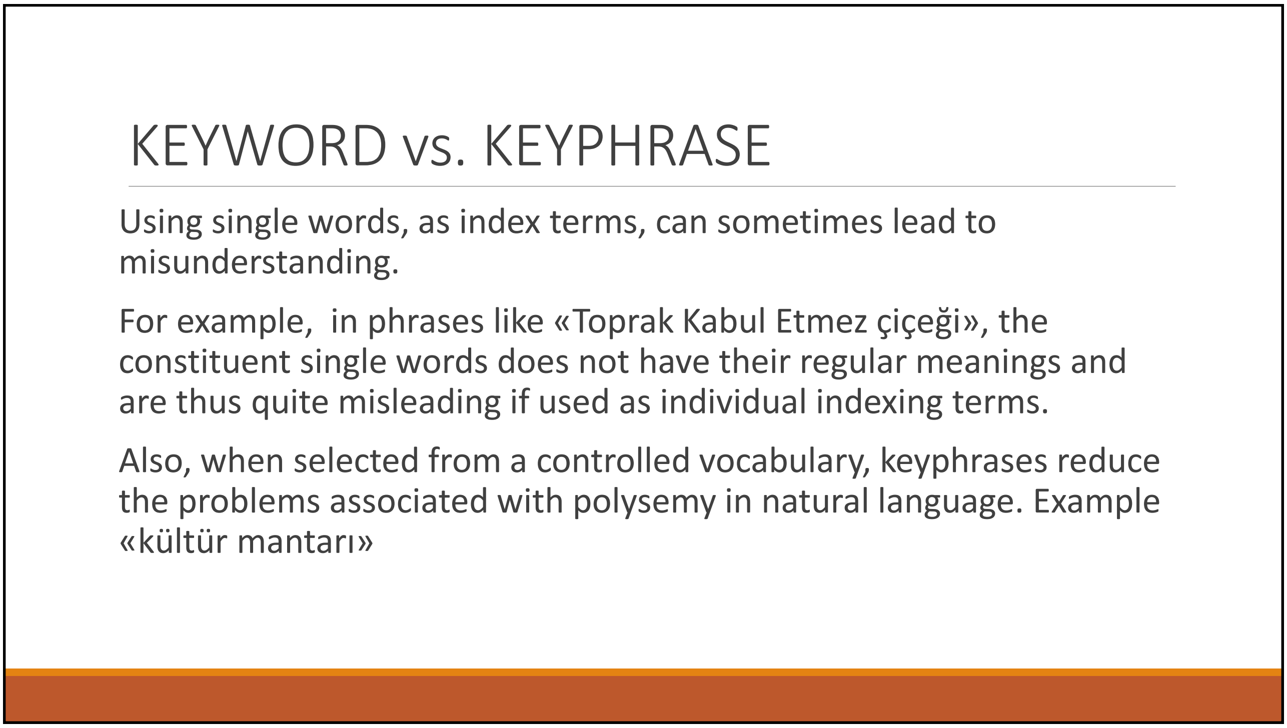
Extracting a small set of units, composed of one or more terms, from a single document is an important problem in Text Mining and Natural Language Processing.

Both single words (keywords) and phrases (key phrases) may be referred to as «key terms».



KEYWORD and KEYPHRASE

- A keyword is single word term.
  - Computer,
  - Disk
- A keyphrase describes a multi-word lexeme.
  - Computer science engineering,
  - Natural language processing



KEYWORD vs. KEYPHRASE

Using single words, as index terms, can sometimes lead to misunderstanding.

For example, in phrases like «Toprak Kabul Etmez çiçeği», the constituent single words does not have their regular meanings and are thus quite misleading if used as individual indexing terms.

Also, when selected from a controlled vocabulary, keyphrases reduce the problems associated with polysemy in natural language. Example «kültür mantarı»

KEYWORD EXTRACTION APPROACHES

There are many approaches by which keyword extraction can be carried out.

- ✓Rule Based Linguistic approaches,
- ✓Statistical approaches
- ✓Machine Learning approaches,

RULE BASED LINGUISTIC APPROACHES

The linguistic approaches

- are generally rule based and are derived from the Linguistic knowledge/features (post tagging: keyterms are noun).
- use the linguistic features of the words, sentences and documents.

STATISTICAL APPROACHES

These approaches are generally based on linguistic corpus and statistical feature derived from the corpus.

Most important advantage of them is that they are **independent of the language** on which they are applied and hence the same technique **can be used on multiple languages**.

These methods may **not give as accurate results** compared to linguistic ones, but the **availability of large amount of datasets** has made it possible to perform statistical analysis and achieve good results.

STATISTICAL APPROACHES

- A Work on Keyword Extraction (Yıldız, 2006)
  - ✓TF-IDF
  - ✓Chi-Square Measure
  - ✓Information Gain
- Keyword Extraction From A Sıngle Document Usıng Word Co-occurrence (Matsuo and Ishızuka, 2003)
  - ✓Co-occurrence + Chi-Square Measure

A WORK ON MULTI-DOCUMENT KEYWORD EXTRACTION (Yıldız, 2006)

“Yıldız Technical University Yıldızlı Hat” is a system that students, staffers and academicians can make suggestions, complaint and get knowledges.

They send their messages by an electronic form. These messages is sent to relevant department by an personnel.

The relevant department makes a get back in the shortest time as soon as possible.

A WORK ON MULTI-DOCUMENT KEYWORD EXTRACTION (Yıldız, 2006)

- Mostly departments gets messages is below.
  - The European Community Office,
  - Directorate of Maintenance And Repair
  - Department of İnformation Technologies
  - Scholarship Office
  - Department Of Library
- If these messages sends directly to the relevant department it would be better.

A WORK ON MULTI-DOCUMENT KEYWORD EXTRACTION (Yıldız, 2006)

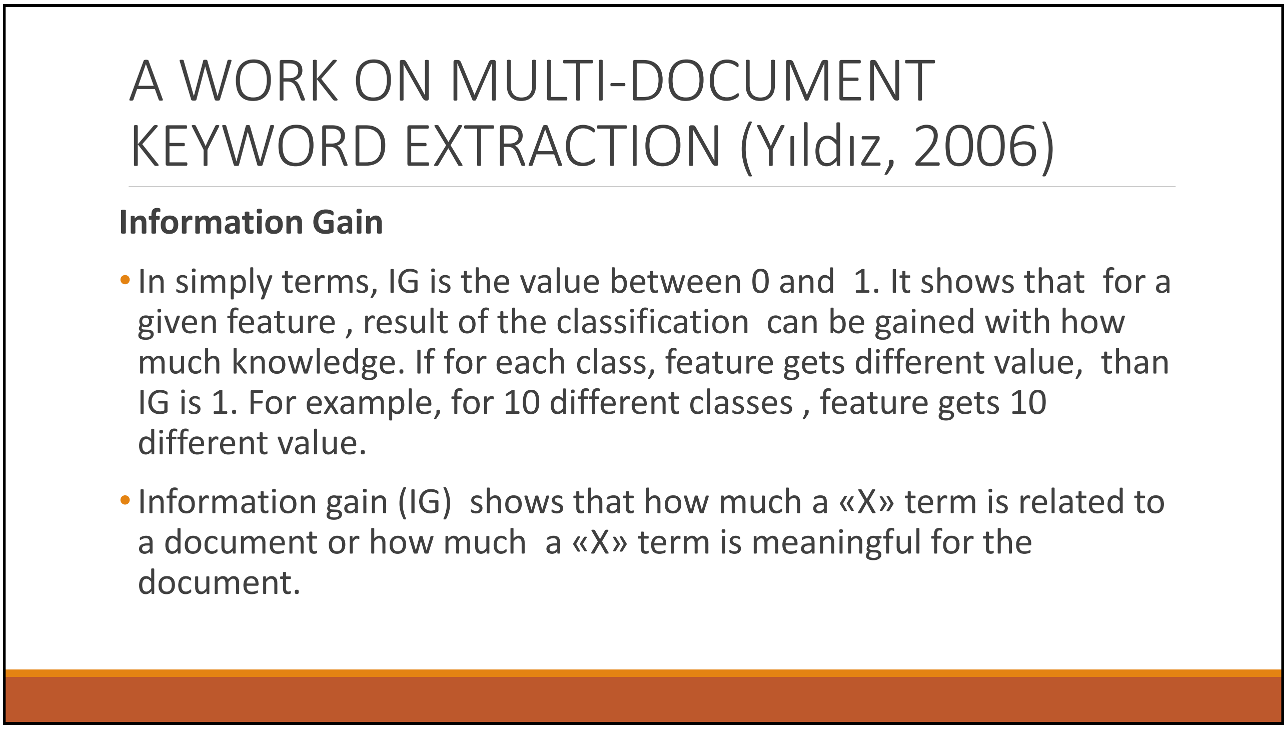
TF-IDF

- TF-IDF does not show just the frequency of the term in the document. It also evaluate the situation of the term in other documents.
- For example, for “The European Community office”, the term “Erasmus” is shown in the messages with high frequency. But in other departments that term may not be shown this much. So, The term “Erasmus” is a keyterm for “The department of European Community”.

A WORK ON MULTI-DOCUMENT KEYWORD EXTRACTION (Yıldız, 2006)

Chi-Square Measure

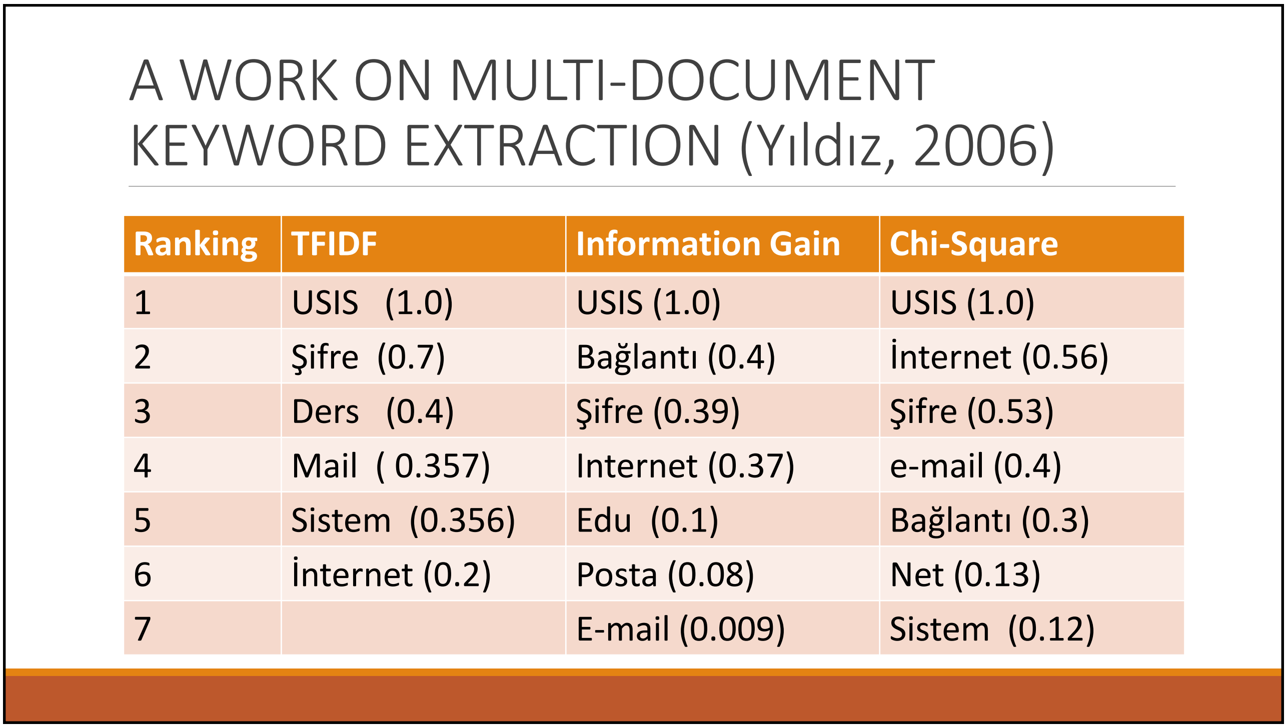
- Chi-Square measure is used for rejecting a hypothesis.
- In keyword extraction, chi-square is found by applying the reverse of the hypothesis. For example, the hypothesis is «term X is not a keyterm». If the chi-square is a big value this means that the hypothesis is rejected and X is a key term.
- Because of low computing cost and easy implementation, it is frequently used in keyterm extraction.



A WORK ON MULTI-DOCUMENT KEYWORD EXTRACTION (Yıldız, 2006)

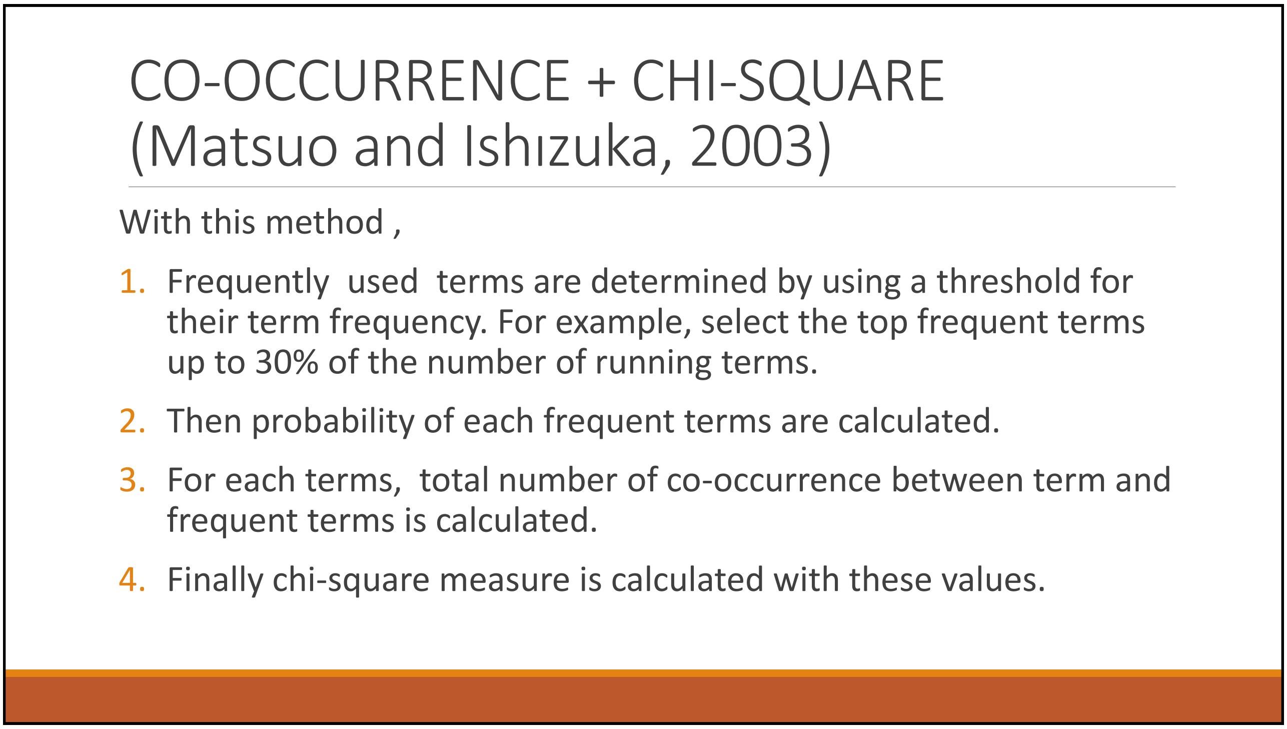
Information Gain

- In simply terms, IG is the value between 0 and 1. It shows that for a given feature , result of the classification can be gained with how much knowledge. If for each class, feature gets different value, than IG is 1. For example, for 10 different classes , feature gets 10 different value.
- Information gain (IG) shows that how much a «X» term is related to a document or how much a «X» term is meaningful for the document.



A WORK ON MULTI-DOCUMENT KEYWORD EXTRACTION (Yıldız, 2006)

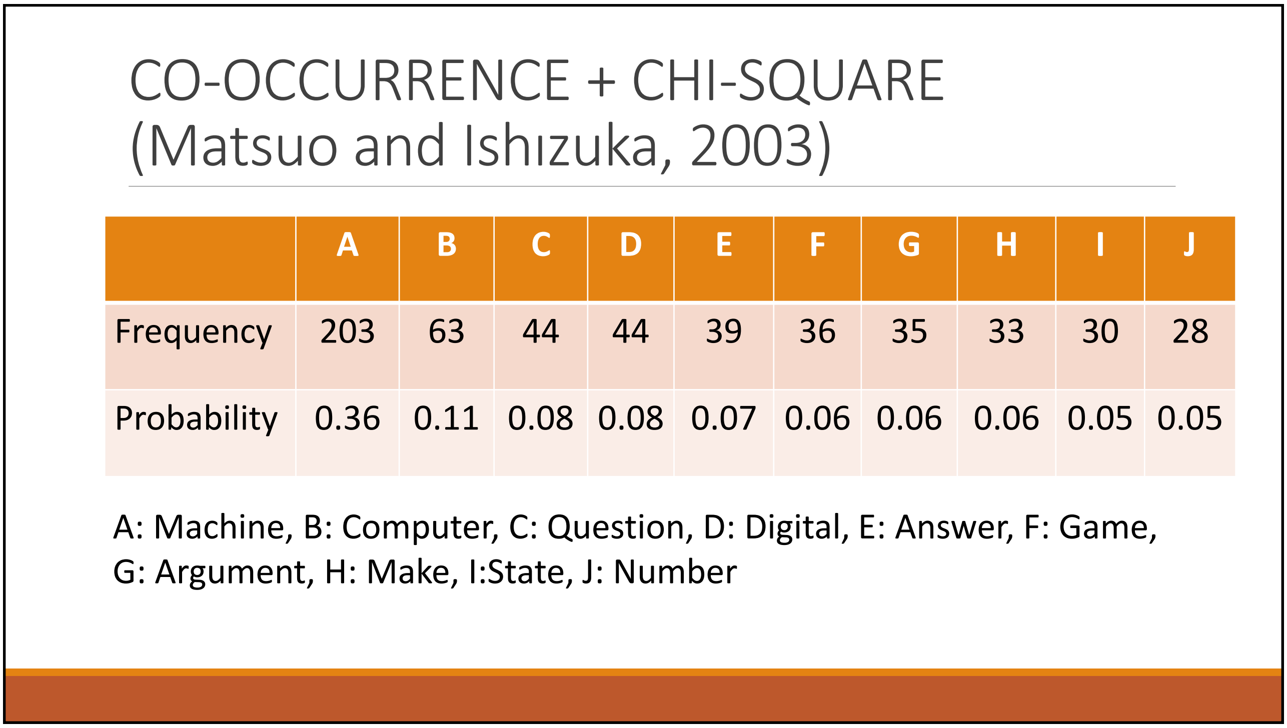
Ranking	TFIDF	Information Gain	Chi-Square
1	USIS (1.0)	USIS (1.0)	USIS (1.0)
2	Şifre (0.7)	Bağlantı (0.4)	İnternet (0.56)
3	Ders (0.4)	Şifre (0.39)	Şifre (0.53)
4	Mail (0.357)	Internet (0.37)	e-mail (0.4)
5	Sistem (0.356)	Edu (0.1)	Bağlantı (0.3)
6	İnternet (0.2)	Posta (0.08)	Net (0.13)
7		E-mail (0.009)	Sistem (0.12)



CO-OCCURRENCE + CHI-SQUARE (Matsuo and Ishızuka, 2003)

With this method ,

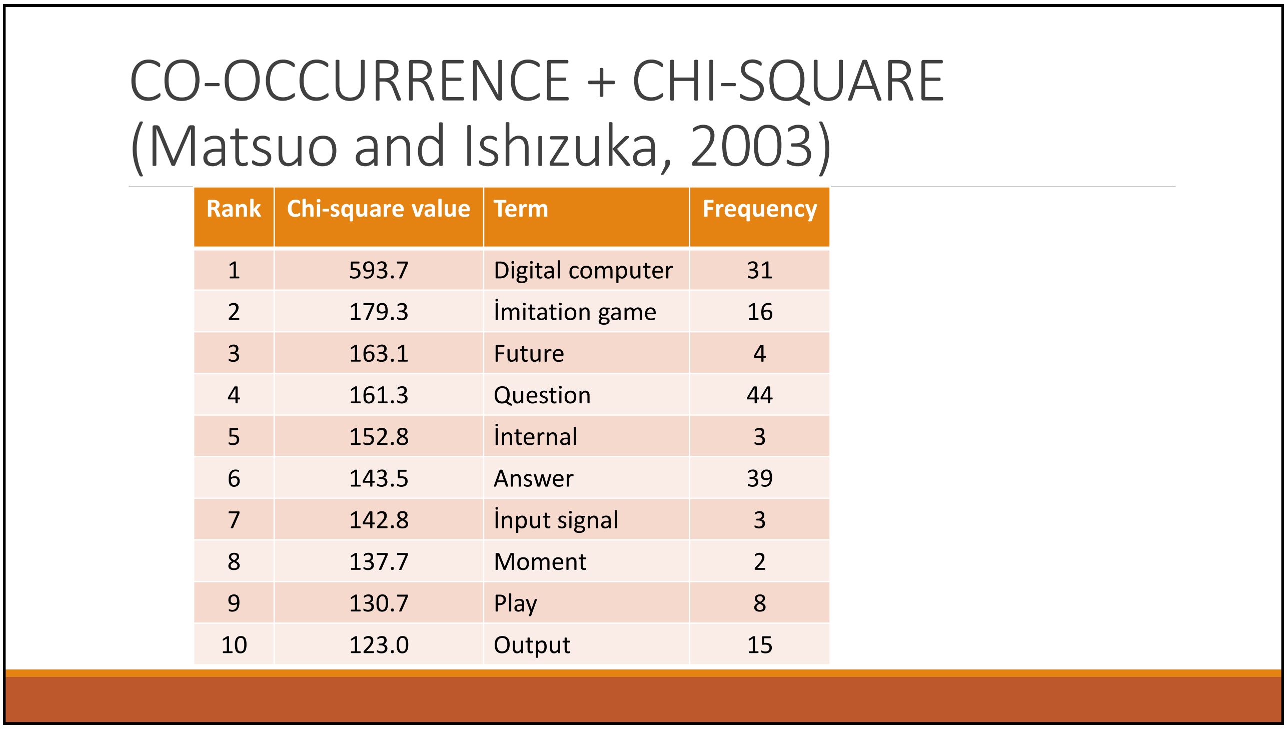
1. Frequently used terms are determined by using a threshold for their term frequency. For example, select the top frequent terms up to 30% of the number of running terms.
2. Then probability of each frequent terms are calculated.
3. For each terms, total number of co-occurrence between term and frequent terms is calculated.
4. Finally chi-square measure is calculated with these values.



CO-OCCURRENCE + CHI-SQUARE (Matsuo and Ishızuka, 2003)

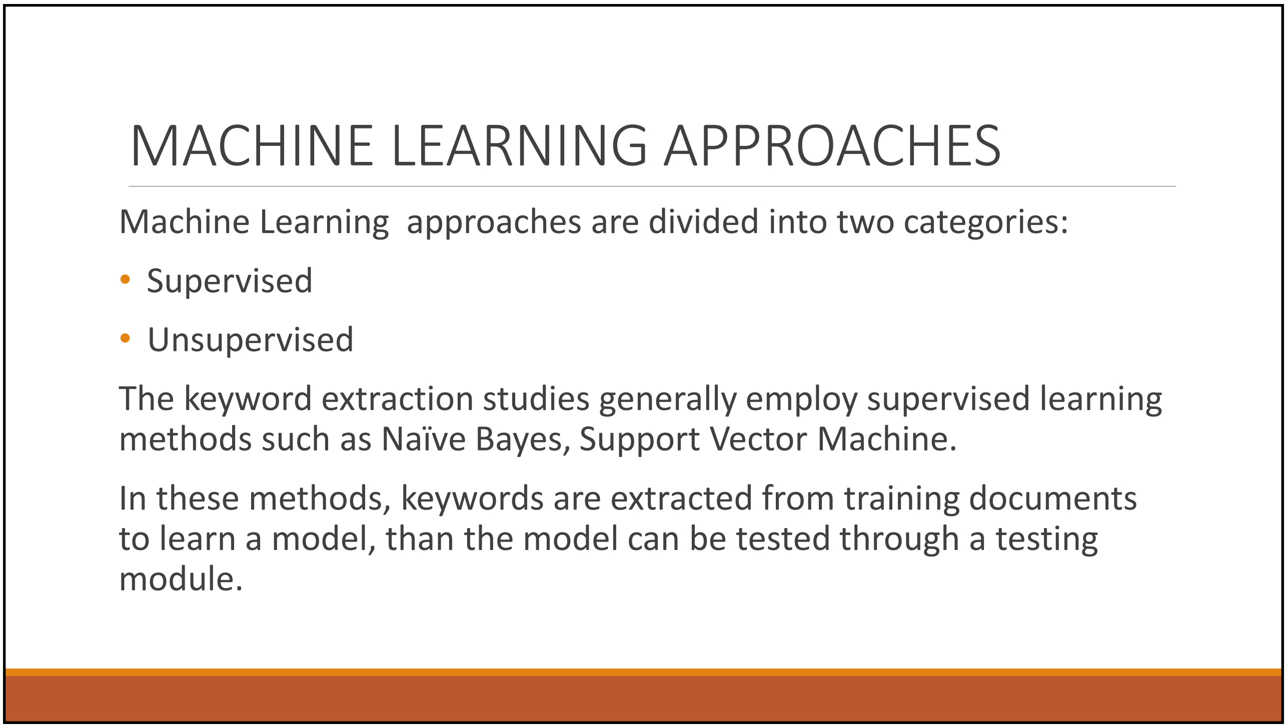
	A	B	C	D	E	F	G	H	I	J
Frequency	203	63	44	44	39	36	35	33	30	28
Probability	0.36	0.11	0.08	0.08	0.07	0.06	0.06	0.06	0.05	0.05

A: Machine, B: Computer, C: Question, D: Digital, E: Answer, F: Game, G: Argument, H: Make, I:State, J: Number



CO-OCCURRENCE + CHI-SQUARE (Matsuo and Ishızuka, 2003)

Rank	Chi-square value	Term	Frequency
1	593.7	Digital computer	31
2	179.3	İmitation game	16
3	163.1	Future	4
4	161.3	Question	44
5	152.8	İnternal	3
6	143.5	Answer	39
7	142.8	İnput signal	3
8	137.7	Moment	2
9	130.7	Play	8
10	123.0	Output	15



MACHINE LEARNING APPROACHES

Machine Learning approaches are divided into two categories:

- Supervised
- Unsupervised

The keyword extraction studies generally employ supervised learning methods such as Naïve Bayes, Support Vector Machine.

In these methods, keywords are extracted from training documents to learn a model, than the model can be tested through a testing module.

MACHINE LEARNING APPROACHES

After a satisfactory model is built, it is used to find keywords from new documents.

However, supervised learning methods require a tagged document corpus which is difficult to build.

In absence of such a corpus, unsupervised or semi-supervised learning methods are used.

DOCUMENT SUMMARIZATION

Ebru Uzundere, Elda Dedja, Banu Diri, M.Fatih Amasyalı, “Türkçe Haber Metinleri İçin Otomatik Özetleme”, ASYU 2008.

ÖZET CÜMLE	METİN CÜMLESİ
Günümüzde internet dünyasının gelişmesi hayal edilemeyen bir bilgi fazlalığını da ortaya çıkarmıştır.	**İçerdiği bilgi miktarının her geçen gün artması ile istenilen bilgiye hızlı erişmek önem kazanmıştır.**
Bunun sonucunda bilgisayarların yardımıyla Otomatik Metin Özetleme sistemleri geliştirilmektedir.	**Otomatik metin özetleme, arama motorlarında ve diğer bazı sistemlerde çok kullanışlı olabilir.**
Özet çıkarma işlemini başarılı bir şekilde gerçekleştirmenin yolu, bu işlemin insanlar tarafından yapılmasıdır, ancak her haberin, makalenin, vd. elle özetlenmesi oldukça zordur.	**? ? ? ? ?**
Bu çalışmada, özeti çıkarılacak metnin cümleleri çeşitli özelliklerine göre puanlanmış, kullanıcının istediği özetleme oranına göre en yüksek puanlı cümleler seçilerek metnin özeti çıkarılmıştır.	**Her cümlenin puanı hesaplandıktan sonra, kullanıcı tarafından verilen özetleme yüzdesine göre, en yüksek puana sahip cümleler özet metinde yer almaktadır.**
Geliştirdiğimiz otomatik haber özetleme sisteminin performansı, sistem ve kullanıcıların özet olarak çıkardığı cümlelerin aynı olma olasılığı alınarak ölçülmüş ve yaklaşık %55 olarak bulunmuştur.	**Çalışmamızda 10 farklı haber metninin 15 kullanıcı tarafından özeti çıkarılmış, daha sonra sistemin verdiği özet metinleri ile karşılaştırması yapılarak başarı yaklaşık %55 olarak elde edilmiştir.**

DOCUMENT SUMMARIZATION

- **Document summarization** is the process of shortening a text document with software, in order to create a summary with the major points of the original document.
- There are two general approaches to document summarization: **extraction** and **abstraction**.

EXTRACTION-BASED SUMMARIZATION

- The automatic system extracts objects from the entire collection, without modifying the objects themselves.
- Examples of this include keyphrase extraction, where the goal is to select individual words or phrases to "tag" a document, and document summarization, where the goal is to select whole sentences (without modifying them) to create a short paragraph summary.
- Similarly, in image collection summarization, the system extracts images from the collection without modifying the images themselves.

ABSTRACTION-BASED SUMMARIZATION

Extraction techniques merely copy the information deemed most important by the system to the summary (for example, key clauses, sentences or paragraphs), while abstraction involves paraphrasing sections of the source document.

The sentence " She hates apple, orange, pear." is summarized as "She hates fruits"

In general, abstraction can condense a text more strongly than extraction, but the programs that can do this are harder to develop as they require use of natural language generation technology, which itself is a growing field.

ABSTRACTION-BASED SUMMARIZATION

While some works have been done in abstractive summarization (creating an abstract synopsis like that of a human), the majority of summarization systems are extractive (selecting a subset of sentences to place in a summary).

SCORING FOR SUMMARIZATION

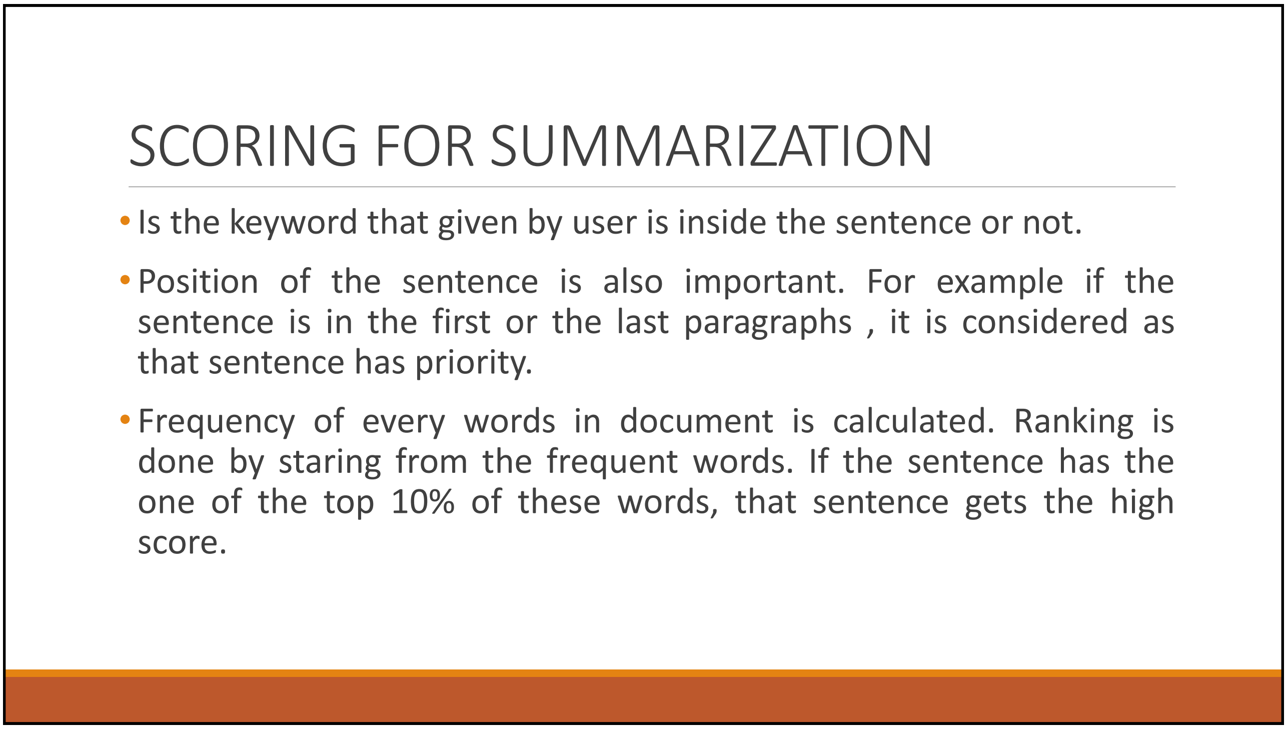
- When the text summarization studies are examined, it is seen that the method of scoring the sentences in the text is shine out.
- To score the sentences, document has to be separated to paragraphs, paragraphs hast to be separated to sentences and sentences has to be separated to words.

SCORING FOR SUMMARIZATION

It is controlled

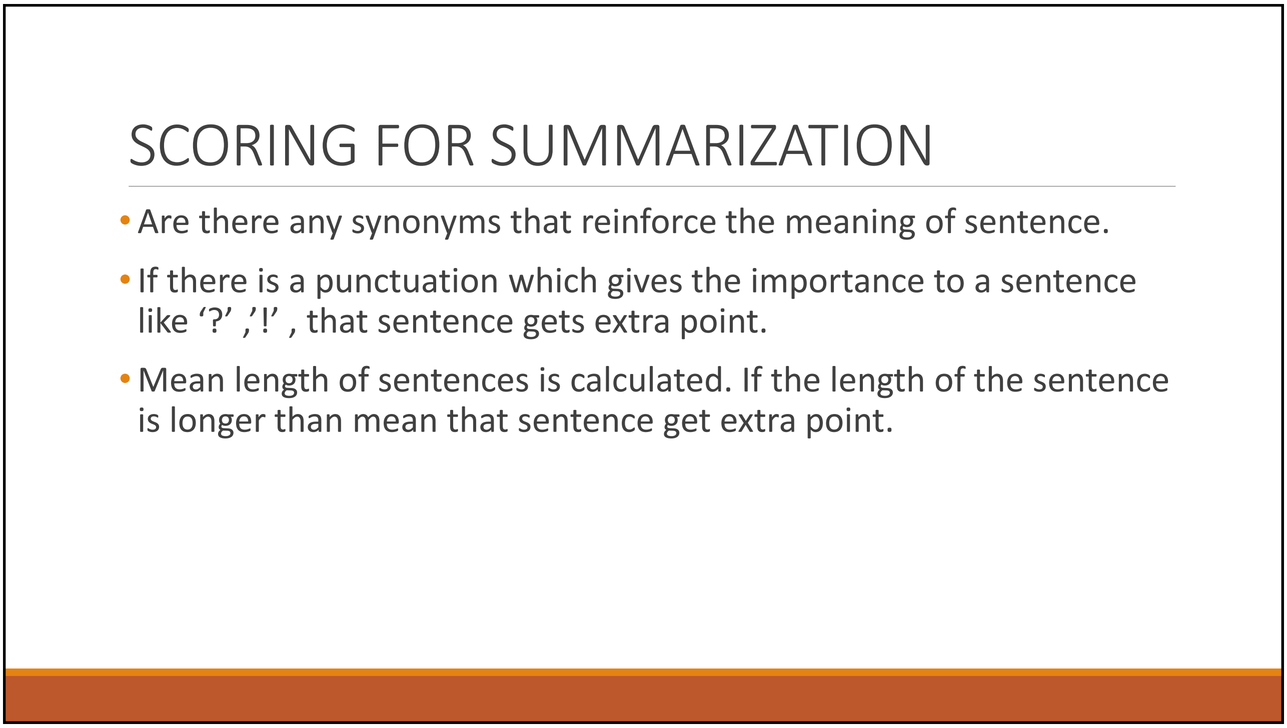
- wether the words in the title of document is inside of the document or not,
- wether the sentence is includes date knowledge or not,
- wether the sentence includes proper noun or not,
- is there any word determined as positive/ negative ,

After **Positive words** («in brief», «last», «as a result») a conclusion sentence is come.
After **Negative words** («but», «because» «such»), detail about the subject is given.



SCORING FOR SUMMARIZATION

- Is the keyword that given by user is inside the sentence or not.
- Position of the sentence is also important. For example if the sentence is in the first or the last paragraphs , it is considered as that sentence has priority.
- Frequency of every words in document is calculated. Ranking is done by staring from the frequent words. If the sentence has the one of the top 10% of these words, that sentence gets the high score.



SCORING FOR SUMMARIZATION

- Are there any synonyms that reinforce the meaning of sentence.
- If there is a punctuation which gives the importance to a sentence like ‘?’ ,’!’ , that sentence gets extra point.
- Mean length of sentences is calculated. If the length of the sentence is longer than mean that sentence get extra point.