Course Outline

- Question Answering systems (Q.A.) +lab (Q.A. finetune BERT model)
- Multilingual and Multimodal Q.A. +lab (Q.A. test BERT models)
- Information Retrieval (I.R.) +lab (I.R. train and test BM25 model)
- Open Domain Q.A. +assingment (Open Domain Q.A)

Course Outline

- **Question Answering systems (Q.A.)** +lab
- **Multilingual and Multimodal Q.A.** +lab
- Information Retrieval (I.R.) +lab
- Open Domain Q.A. +assignment

summary!

What is question answering?

The goal of question answering is to build systems that automatically answer questions posed by humans in a natural language

Q.A. is a Reading Comprehension problem

Reading comprehension = comprehend a passage of text and answer questions about its content (P, Q) $\longrightarrow$ A

Tesla was the fourth of five children. He had an older brother named Dane and three sisters, Milka, Angelina and Marica. Dane was killed in a horse-riding accident when Nikola was five. In 1861, Tesla attended the "Lower" or "Primary" School in Smiljan where he studied German, arithmetic, and religion. In 1862, the Tesla family moved to Gospić, Austrian Empire, where Tesla's father worked as a pastor. Nikola completed "Lower" or "Primary" School, followed by the "Lower Real Gymnasium" or "Normal School."

Q: What language did Tesla study while in school?

A: German

Q.A. is a Reading Comprehension problem

Reading comprehension = comprehend a passage of text and answer questions about its content (P, Q) $\longrightarrow$ A

Tesla was the fourth of five children. He had an older brother named Dane and three sisters, Milka, Angelina and Marica. Dane was killed in a horse-riding accident when Nikola was five. In 1861, Tesla attended the "Lower" or "Primary" School in Smiljan where he studied German, arithmetic, and religion. In 1862, the Tesla family moved to Gospić, Austrian Empire, where Tesla's father worked as a pastor. Nikola completed "Lower" or "Primary" School, followed by the "Lower Real Gymnasium" or "Normal School."

Q: What language did Tesla study while in school?

A: German

Extractive Q.A.

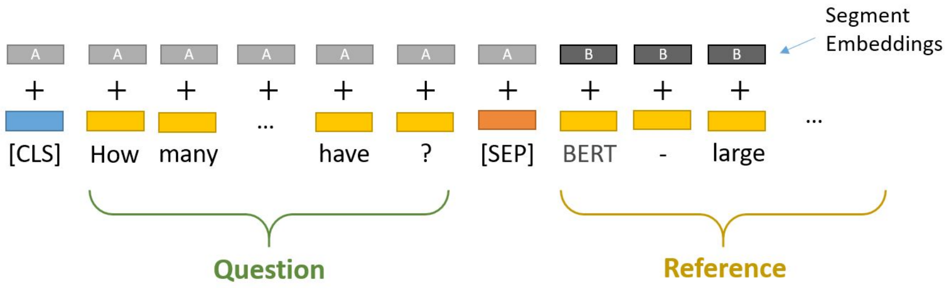
Models for reading comprehension

- **How can we build a model to solve SQuAD?**
  - Problem formulation (extractive Q.A.)
    - Input:

$$C = (c_1, c_2, \ldots, c_N),\ Q = (q_1, q_2, \ldots, q_M),\ c_i, q_i \in V$$

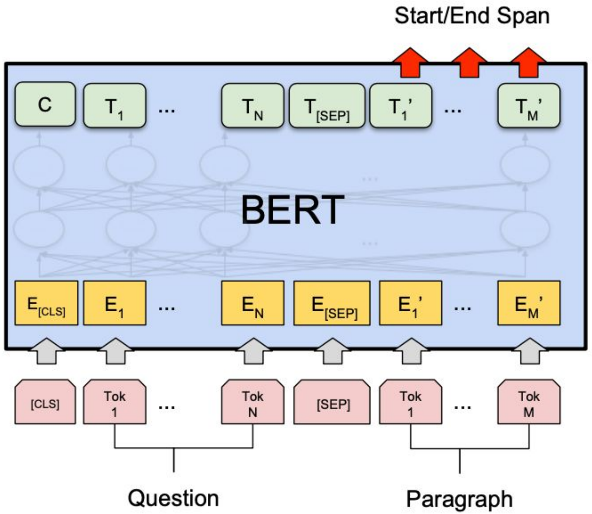
      - c: passage or document
      - q: question or query
    - Output: 1 ≤ start ≤ end ≤ N (answer is a spam in the passage)
- **From pipelined models ~ LSTM based ~ BERT like models**

BERT based Neural Models for reading comprehension



Question: `How many parameters does BERT-large have?`

Reference Text: `BERT-large is really big... it has 24 layers and an embedding size of 1,024, for a total of 340M parameters! Altogether it is 1.34GB, so expect it to take a couple minutes to download to your Colab instance.`



- **Input:**
  - tokenization (BPE)
  - segment ids
  - CLS & SEP
- **Output:**
  - Argmax start & end idx.
  - Recover answer (BPE)
  - Get score

Multilingual Neural Models for reading comprehension

Problem: most of the datasets in NLP are in English… SQuAD, QuAC….

- Translate Train (MT + alingment)
- Translate Test (MT + alingment)
- **Zero-Shot** (CrossLingual pretrained L.M.)

SQuAD Train (en) | **SQuAD dev (en)** → **Test in X**

We don’t need M.T. and alignment!

Is reading comprehension solved?

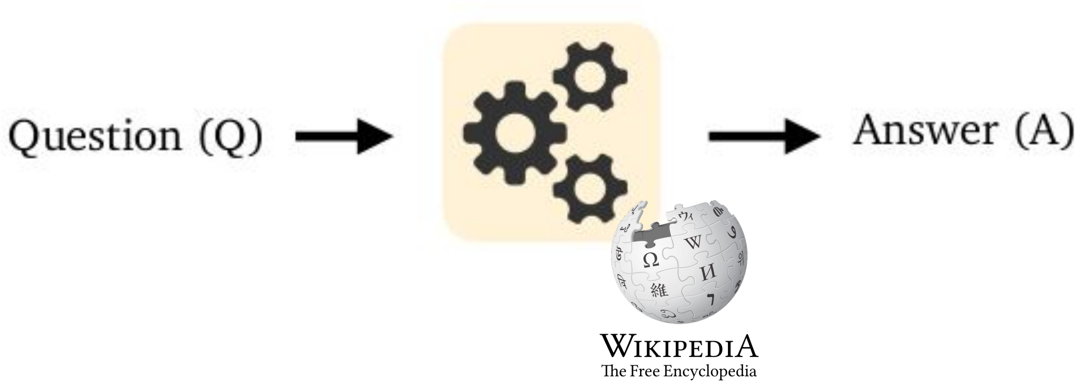
- **Of course not!**
- Super human performance on SQuAD
- Out of domain Q.A. is still a problem
- Extractive models -> **We assume allways a given passage**

Is reading comprehension solved?

- **Of course not!**
- Super human performance on SQuAD
- Out of domain Q.A. is still a problem
- Extractive models -> **We assume allways a given passage**

Open Domain Q.A.

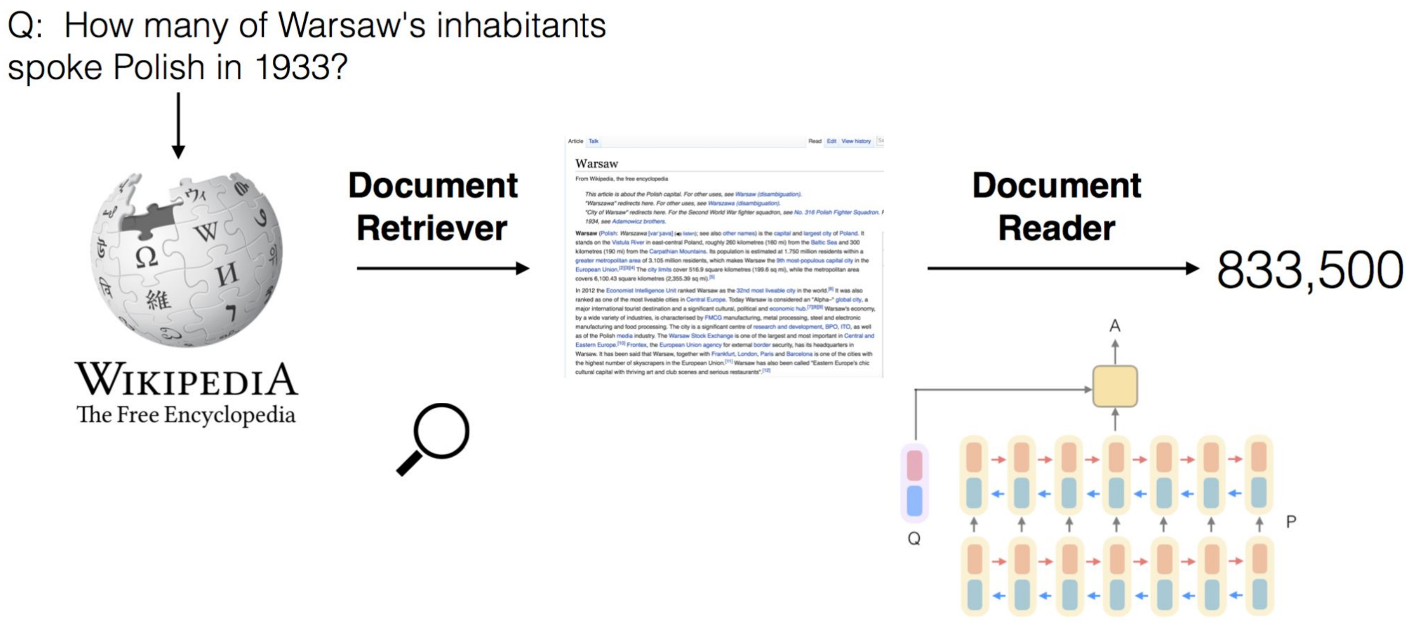
Open Domain Q.A.



- Different from reading comprehension, we don’t assume a given passage.
- Instead, we only have access to a large collection of documents (e.g., Wikipedia). We don’t know where the answer is located, and the goal is to return the answer for any open-domain questions.
- Much more challenging but **a more practical problem!**

Open Domain Q.A.

Open-domain QA
SQuAD, TREC, WebQuestions, WikiMovies



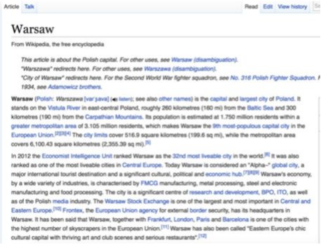
Open Domain Q.A.

Open-domain QA
SQuAD, TREC, WebQuestions, WikiMovies

Q: How many of Warsaw's inhabitants spoke Polish in 1933?

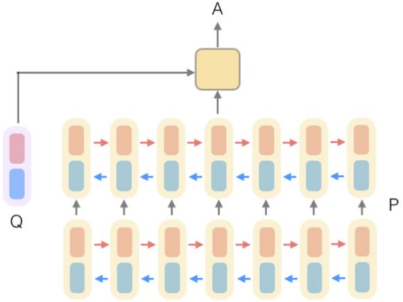
WIKIPEDIA
The Free Encyclopedia

Document Retriever



Document Reader

833,500



Information Retrieval Model (I.R.)

BERT
fine-tuned for Q.A.

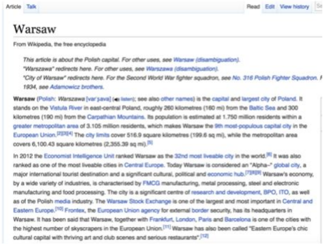
Open Domain Q.A.

Open-domain QA
SQuAD, TREC, WebQuestions, WikiMovies

This is your assignment :D

Q: How many of Warsaw's inhabitants spoke Polish in 1933?

Document Retriever


Article Talk Read Edit View history

Warsaw

From Wikipedia, the free encyclopedia

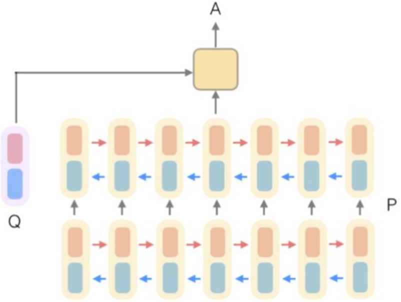
This article is about the Polish capital. For other uses, see Warsaw (disambiguation).
"Warszawa" redirects here. For other uses, see Warszawa (disambiguation).
"City of Warsaw" redirects here. For the Second World War fighter squadron, see No. 316 Polish Fighter Squadron. For 1934, see Adamowicz brothers.

Warsaw (Polish: *Warszawa* [varˈʂava] (listen); see also other names) is the capital and largest city of Poland. It stands on the Vistula River in east-central Poland, roughly 260 kilometres (160 mi) from the Baltic Sea and 300 kilometres (190 mi) from the Carpathian Mountains. Its population is estimated at 1.750 million residents within a greater metropolitan area of 3.105 million residents, which makes Warsaw the 9th most-populous capital city in the European Union.[2][3][4] The city limits cover 516.9 square kilometres (199.6 sq mi), while the metropolitan area covers 6,100.43 square kilometres (2,355.39 sq mi).[5]

In 2012 the Economist Intelligence Unit ranked Warsaw as the 32nd most liveable city in the world.[6] It was also ranked as one of the most liveable cities in Central Europe. Today Warsaw is considered an "Alpha–" global city, a major international tourist destination and a significant cultural, political and economic hub.[7][8][9] Warsaw's economy, by a wide variety of industries, is characterised by FMCG manufacturing, metal processing, steel and electronic manufacturing and food processing. The city is a significant centre of research and development, BPO, ITO, as well as of the Polish media industry. The Warsaw Stock Exchange is one of the largest and most important in Central and Eastern Europe.[10] Frontex, the European Union agency for external border security, has its headquarters in Warsaw. It has been said that Warsaw, together with Frankfurt, London, Paris and Barcelona is one of the cities with the highest number of skyscrapers in the European Union.[11] Warsaw has also been called "Eastern Europe's chic cultural capital with thriving art and club scenes and serious restaurants".[12]

Document Reader

833,500



BERT
fine-tuned for Q.A.

Information Retrieval Model (I.R.)

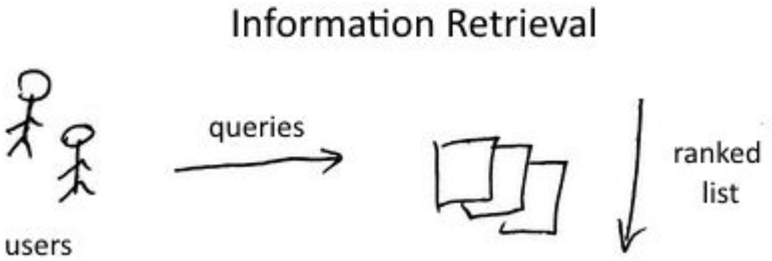
Course Outline

- Question Answering systems (Q.A.) +lab
- Multilingual and Multimodal Q.A. +lab
- **Information Retrieval (I.R.)** +lab
- Open Domain Q.A. +assignment

Course Outline

- **Information retrieval**
  - **Introduction**
  - I.R. practical applications
  - I.R. for text retrieval

What is information retrieval?

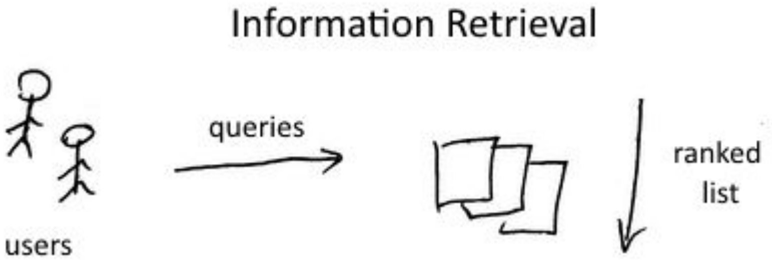


Information retrieval (**IR**) is the process of obtaining information system resources that are relevant to an information need from a collection of those **resources**.

- Full-text
- Books
- Passages
- Music
- ...

image source: http://mlwiki.org/index.php/Information_Retrieval

What is information retrieval?



Automated information retrieval systems are used to reduce what has been called information overload (**Infoxication**).

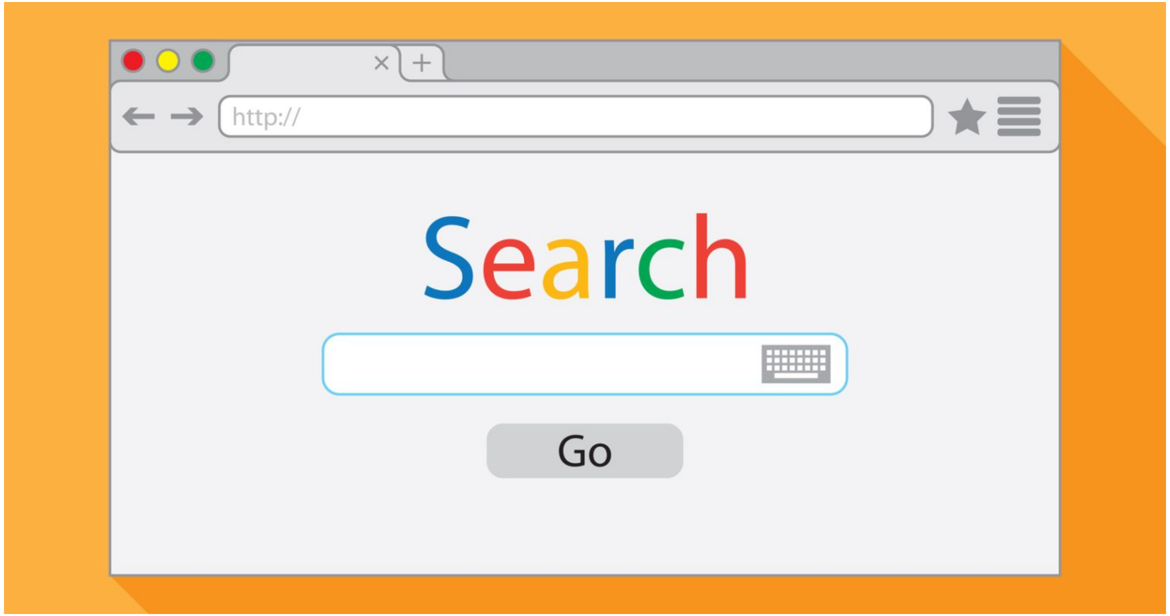
- **Information overload:** is the difficulty in understanding an issue and effectively making decisions when one has too much information about that issue

image source: http://mlwiki.org/index.php/Information_Retrieval

Course Outline

- **Information retrieval**
  - Introduction
  - **I.R. practical applications**
  - I.R. for text retrieval

I.R.: Lots of practical application



https://www.searchenginejournal.com/alternative-search-engines/271409/

Google
Ask
bing
YAHOO!
Aol.

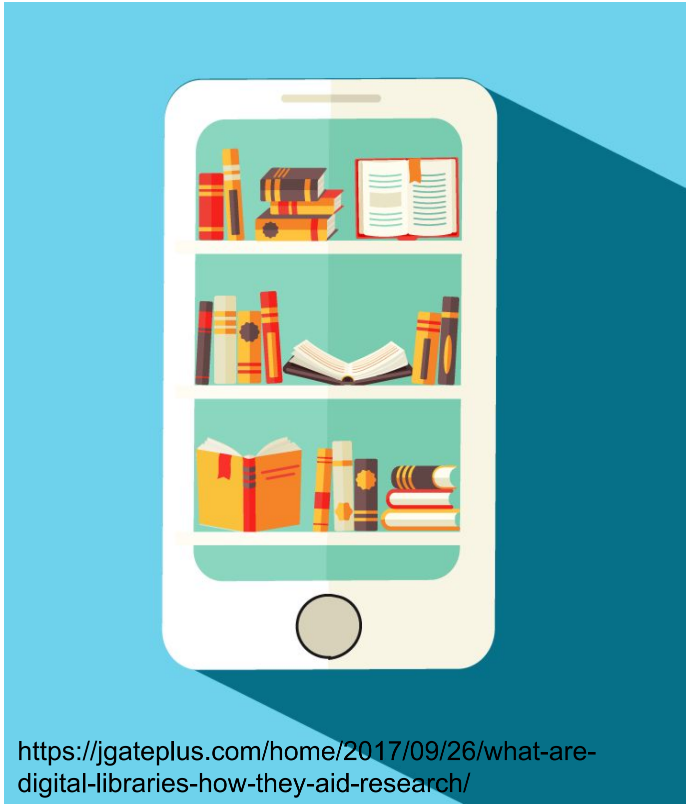
Search engines
Web Search

I.R.: Lots of practical application

https://www.mentionlytics.com/blog/10-best-social-search-engines/

Search engines
Desktop search
Social media search
...

I.R.: Lots of practical application



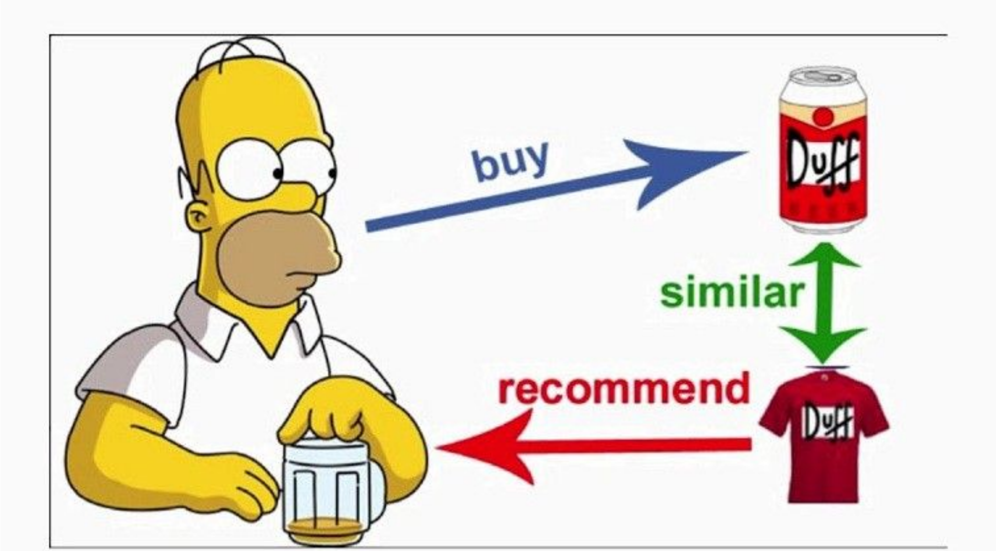
https://jgateplus.com/home/2017/09/26/what-are-digital-libraries-how-they-aid-research/

Also called an **online library**, an **internet library**, a **digital repository**, or a **digital collection** is an online database of digital objects that can include text, still images, audio, video, digital documents, or other digital media formats or a library accessible through the internet.

Digital Libraries

I.R.: Lots of practical application

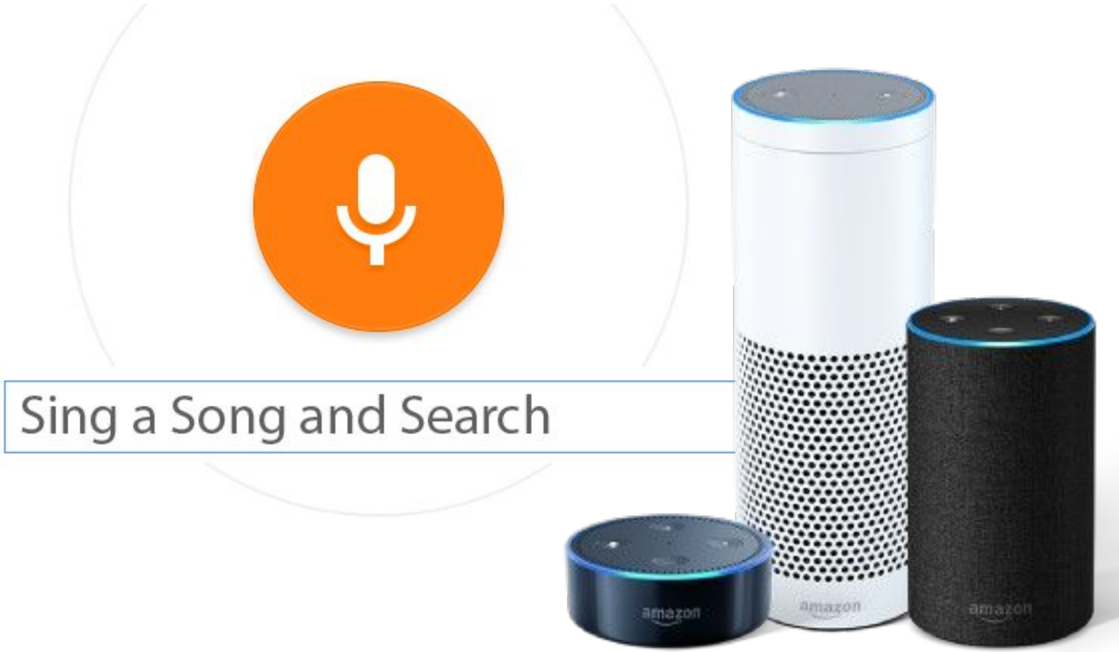
A recommender system, or a recommendation system (sometimes replacing 'system' with a synonym such as platform or engine), is a subclass of information filtering system that seeks to predict the "rating" or "preference" a user would give to an item.



Information Filtering
&
Recomender Systems

https://towardsdatascience.com/build-your-own-recommender-system-within-5-minutes-30dd40388fbf

I.R.: Lots of practical application



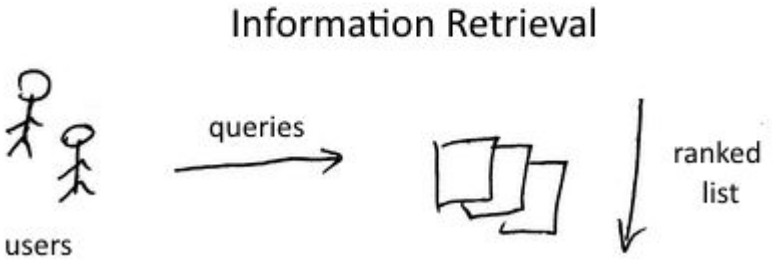
Media search
blogs
images
music
news
speech
video
...

https://towardsdatascience.com/build-your-own-recommender-system-within-5-minutes-30dd40388fbf

Course Outline

- **Information retrieval**
  - Introduction
  - I.R. practical applications
  - **I.R. for text retrieval**

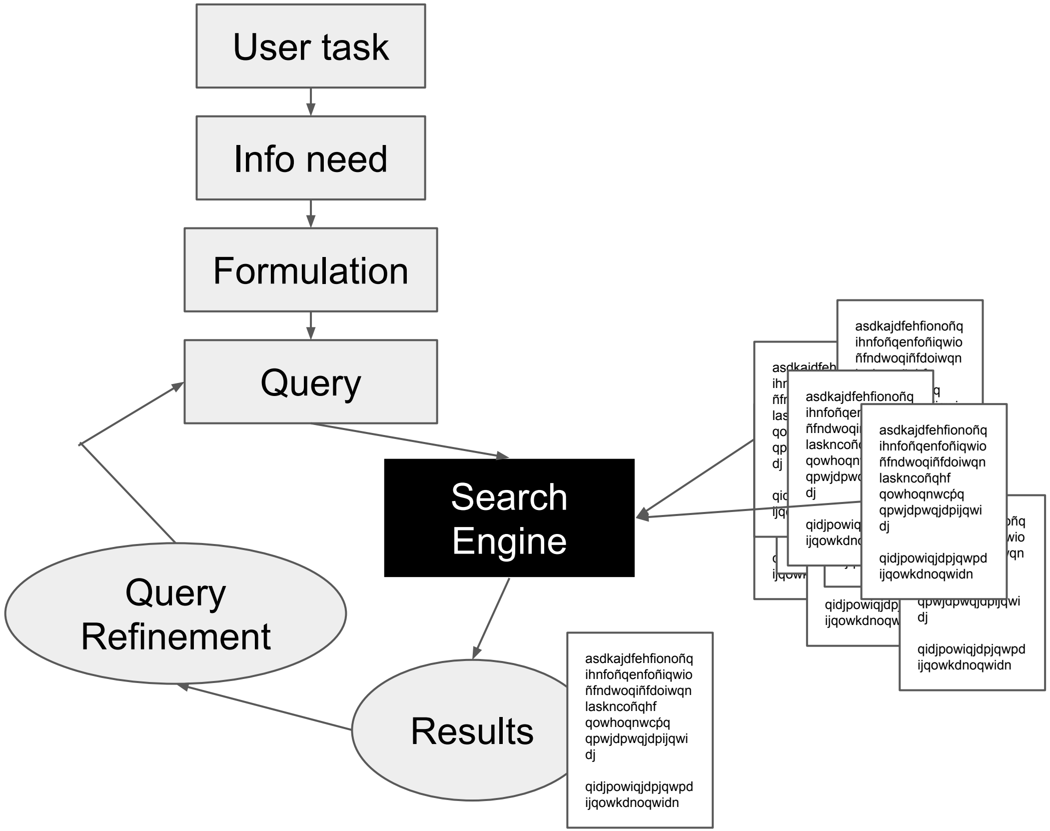
Basic assumption of I.R.



- Collection: A set of documents
  - A static set of unstructured or semistructured documents
- Goal
  - Retrieve documents with information that is **relevant** to the user's **information need** and helps the user **complete a task**

image source: http://mlwiki.org/index.php/Information_Retrieval

The classic search model



Get ride of mice in a politically correct way

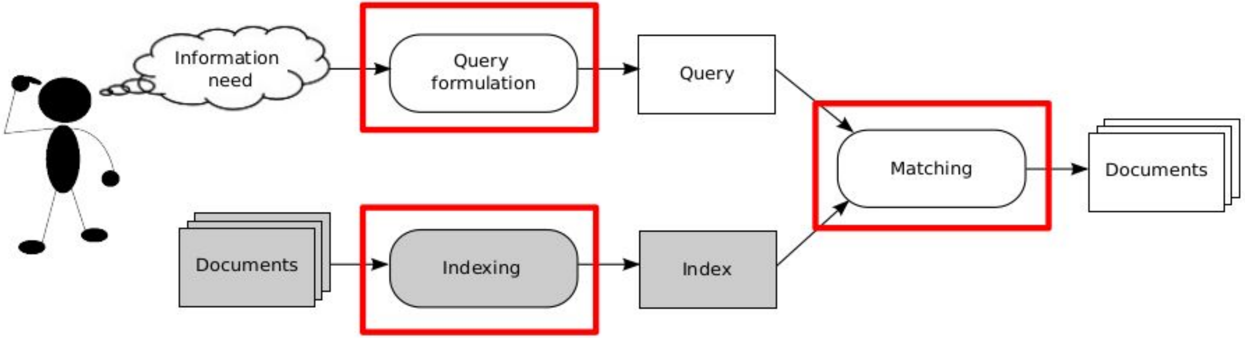
Info about removing mice without killing them

Whats the best way to trap mice alive?

Trap mice alive

The classic search model

- Collecting documents
- **Indexing** documents (representation)
- **Query** formulation: the user formulates a query according to his information need
- **Matching**: the query representation is compared against the document representation (index) to select relevant documents

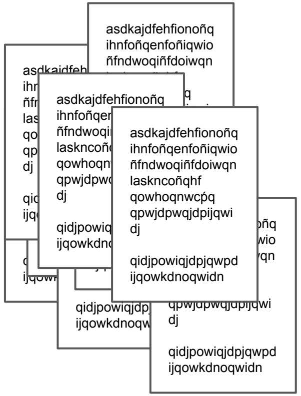


Course Outline

- **Information retrieval**
  - Introduction
  - I.R. practical applications
  - **I.R. for text retrieval**
    - **Boolean retrieval**
    - Ranked retrieval
    - I.R. models

Boolean retrieval

- Term-document incidence matrix
- Inverted index
- Query processing
- The Merge
- Boolean queries



Doc collection: Works of William Shakespeare
Query: Which plays of Shakespeare contain the words Brutus and Caesar, but not Calpurnia?

Boolean retrieval

Doc collection: Works of William Shakespeare
Query: Which plays of Shakespeare contain the words Brutus and Caesar, but not Calpurnia?

- **Term-document incidence matrix**
- Inverted index
- Query processing
- The Merge
- Boolean queries

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	1	1	0	0	0	1
Brutus	1	1	0	1	0	0
Caesar	1	1	0	1	1	1
Calpurnia	0	1	0	0	0	0
Cleopatra	1	0	0	0	0	0
mercy	1	0	1	1	1	1
worser	1	0	1	1	1	0

1 if doc contains word, 0 otherwise

Boolean retrieval

Doc collection: Works of William Shakespeare
Query: Which plays of Shakespeare contain the words Brutus and Caesar, but not Calpurnia?

- **Term-document incidence matrix**
- Inverted index
- Query processing
- The Merge
- Boolean queries

- So we have a 0/1 vector for each term
- **To answer the query:** Brutus AND Caesar BUT NOT Calpurnia
  - Take the vectors for Brutus, Caesar and Calpurnia
  - Bitwise AND

110100 **AND** 110111 **AND** 101111 = **100100**

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	1	1	0	0	0	1
Brutus	1	1	0	1	0	0
Caesar	1	1	0	1	1	1
Calpurnia	0	1	0	0	0	0
Cleopatra	1	0	0	0	0	0
mercy	1	0	1	1	1	1
worser	1	0	1	1	1	0

Boolean retrieval

Doc collection: Works of William Shakespeare
Query: Which plays of Shakespeare contain the words Brutus and Caesar, but not Calpurnia?

- **Term-document incidence matrix**
- Inverted index
- Query processing
- The Merge
- Boolean queries

- So we have a 0/1 vector for each term
- **To answer the query:** Brutus AND Caesar BUT NOT Calpurnia
  - Take the vectors for Brutus, Caesar and Calpurnia
  - Bitwise AND

110100 **AND** 110111 **AND** 101111 = **100100**

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	1	1	0	0	0	1
Brutus	1	1	0	1	0	0
Caesar	1	1	0	1	1	1
Calpurnia	0	1	0	0	0	0
Cleopatra	1	0	0	0	0	0
mercy	1	0	1	1	1	1
worser	1	0	1	1	1	0

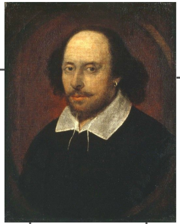
Boolean retrieval

Doc collection: Works of William Shakespeare
Query: Which plays of Shakespeare contain the words Brutus and Caesar, but not Calpurnia?

- **Term-document incidence matrix**
- Inverted index
- Query processing
- The Merge
- Boolean queries

- So we have a 0/1 vector for each term
- **To answer the query:** Brutus AND Caesar BUT NOT Calpurnia
  - Take the vectors for Brutus, Caesar and Calpurnia
  - Bitwise AND

110100 **AND** 110111 **AND** 101111 = **100100**



Antony and Cleopatra, Act III, Scene ii
Agrippa [Aside to DOMITIUS ENOBARBUS]:
Why, Enobarbus, When Antony found Julius **Caesar** dead, He cried almost to roaring; and he wept When at Philippi he found **Brutus** slain.

Hamlet, Act III, Scene ii
Lord Polonius: I did enact Julius **Caesar**
I was killed i' the Capitol;
Brutus killed me.

Boolean retrieval

Doc collection: Works of William Shakespeare
Query: Which plays of Shakespeare contain the words Brutus and Caesar, but not Calpurnia?

- **Term-document incidence matrix**
- Inverted index
- Query processing
- The Merge
- Boolean queries

- Consider 1 million documents, each with about 1000 words
- Avg 6 bytes/word including spaces/punctuation
- 6GB of data in the documents
- Say there are 500K distinct terms among these
- 500K x 1M matrix has half-a-trillion 0's and 1's
  Too big!!
- But it has no more than one billion 1's
  - Matrix is extremely sparse
  - A minimum of 99.8% of the cells are 0
- What's a better representation?
  - **We only record the 1 positions**

Boolean retrieval

Doc collection: Works of William Shakespeare
Query: Which plays of Shakespeare contain the words Brutus and Caesar, but not Calpurnia?

- **Term-document incidence matrix**
- **Inverted index**
- Query processing
- The Merge
- Boolean queries

- For each term t, we must store a list of all documents that contain t
  - Identify each document by a docID
- Can we used fixed-size arrays for this?
- We need variable-size posting lists
  - On disk, a continuous run of postings is normal and best
  - In memory, can use linked list or variable length arrays

Brutus	1 2 4 11 31 45 173 174
Caesar	1 2 4 5 6 16 57 132
Calpurnia	2 31 54 101

Boolean retrieval

Doc collection: Works of William Shakespeare
Query: Which plays of Shakespeare contain the words Brutus and Caesar, but not Calpurnia?

- **Term-document incidence matrix**
- **Inverted index**
- **Query processing**
- **The Merge**
- Boolean queries

- Consider processing the query: **Brutus AND Caesar**
- Locate Brutus in the dictionary
- Locate Caesar in the dictionary
- “Merge” **intersect the document sets**
- Retrieve the documents

Brutus	**1** 2 **4** 11 31 45 173 174
Caesar	**1** 2 **4** 5 6 16 57 132
Calpurnia	2 31 54 101

Boolean retrieval

Doc collection: Works of William Shakespeare
Query: Which plays of Shakespeare contain the words Brutus and Caesar, but not Calpurnia?

- **Term-document incidence matrix**
- **Inverted index**
- **Query processing**
- **The Merge**
- **Boolean queries**

- The boolean retrieval model is being able to ask a query that is a boolean expression
  - Boolean queries are queries using AND, OR and NOT to join query terms
  - Perhaps the simplest model to build an IR system on
- Primary commercial retrieval tool for 3 decades
- Many search system you still use are boolean
  - Email, library catalog

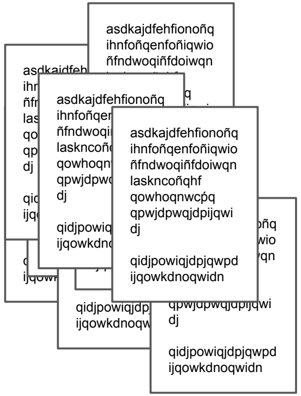
Brutus	**1** 2 **4** 11 31 45 173 174
Caesar	**1** 2 **4** 5 6 16 57 132
Calpurnia	2 31 54 101

Course Outline

- **Information retrieval**
  - Introduction
  - I.R. practical applications
  - **I.R. for text retrieval**
    - Boolean retrieval
    - **Ranked retrieval**
    - I.R. models

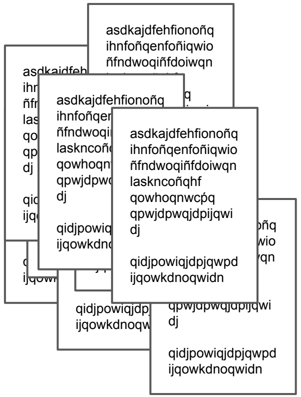
Ranked retrieval Motivation

- Thus far, our queries have been **Boolean**
    - Documents either match or don't
- Good for expert users with precise understanding of their needs and of the collection
- Also good for applications: applications can easily consume 1000s of results
- **Not good for the majority of users**
    - Most users incapable of writing Boolean queries (or they are, but they think it's too much work)
    - Most users don't want to wade trough 1000s of results
        - This is particularly true of web search

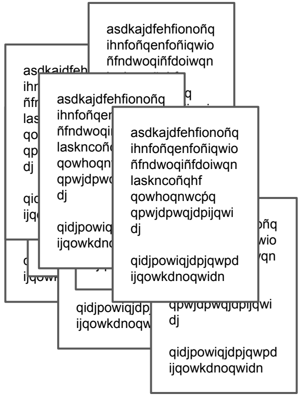


Ranked retrieval Motivation

- Rather than a set of documents satisfying a query expression, the system returns and **ordering** over the documents in the collection for a query
- **Too few or too many:** is not a problem in ranked retrieval
- With ranking, large result sets are not an issue
  - Just show the top k (≈10) results
  - Doesn't overwhelm the user
- **The ranking algorithm works:** More relevant results are ranked higher than less relevant results



Ranked retrieval



Ranked retrieval

- Bag of words model
- Term-document count matrix
- Term frequency
- Document frequency
- IDF
- TF-IDF

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	5.25	3.18	0	0	0	0.35
Brutus	1.21	6.1	0	1	0	0
Caesar	8.59	2.54	0	1.51	0.25	0
Calpurnia	0	1.54	0	0	0	0
Cleopatra	2.85	0	0	0	0	0
mercy	1.51	0	1.9	0.12	5.25	0.88
worser	1.37	0	0.11	4.15	0.25	1.95

Course Outline

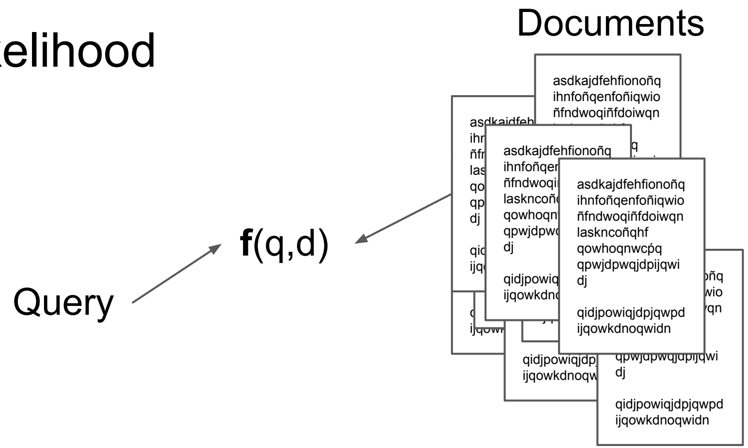
- **Information retrieval**
  - Introduction
  - I.R. practical applications
  - **I.R. for text retrieval**
    - Boolean retrieval
    - Ranked retrieval
    - **I.R. models (Okapi BM25)**

I.R. models

Key challenge: how to measure the likelihood that document d is relevant to query q

Ranking function: f(q,d)

- Similarity-based
- Probabilistic models
  - **Okapi BM25**

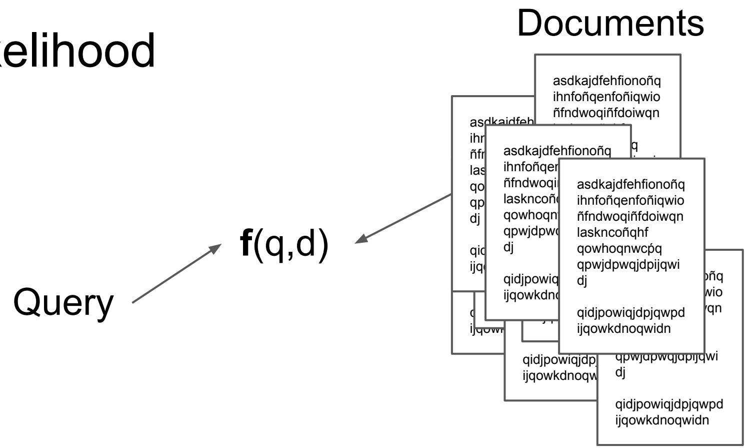


I.R. models **Okapi BM25** (BM = best matching)

Key challenge: how to measure the likelihood that document d is relevant to query q

Ranking function: f(q,d)

- Similarity-based
- Probabilistic models
    - **Okapi BM25**



$$w_{tD} = \frac{(k_1+1)\, tf_{tD}}{k_1\left((1-b)+b\frac{l_D}{avl}\right)+tf_{tD}}\, idf_t$$

For each word in query, compute for each document in the collection
f(q,d)=sum WtD

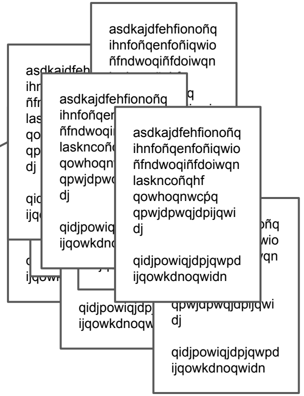
I.R. models **Okapi BM25** (BM = best matching)

Key challenge: how to measure the likelihood that document d is relevant to query q

Ranking function: f(q,d)

- Similarity-based
- Probabilistic models
  - **Okapi BM25** (Robertson and Zaragoza, 2009)

Query → **f**(q,d) ← Documents



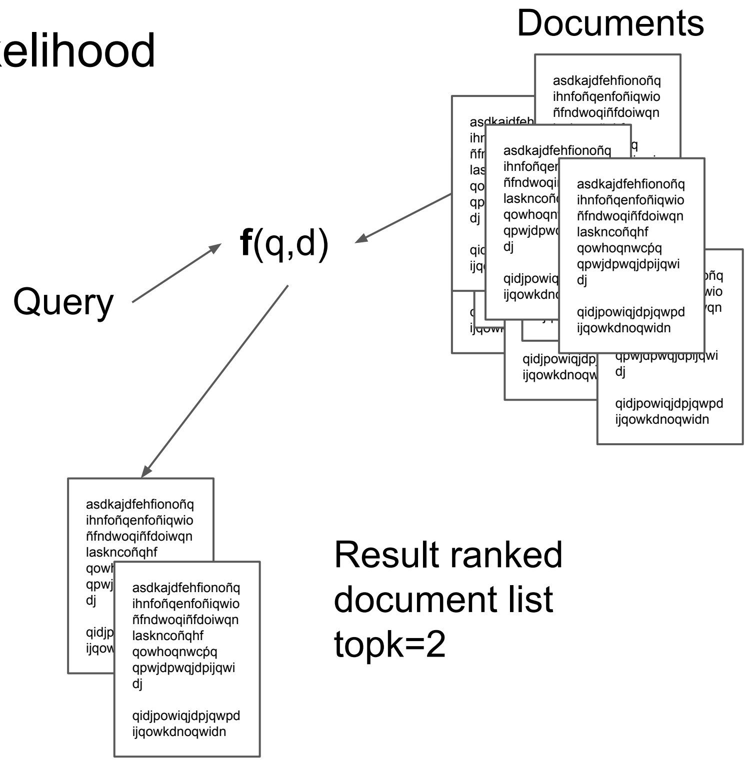
$$w_{tD} = \frac{(k_1+1)\,tf_{tD}}{k_1\left((1-b)+b\frac{l_D}{avl}\right)+tf_{tD}}\,idf_t$$

term frequency of t in D
Documents length
Average document length
K,B Tuning parameters

I.R. models **Okapi BM25** (BM = best matching)

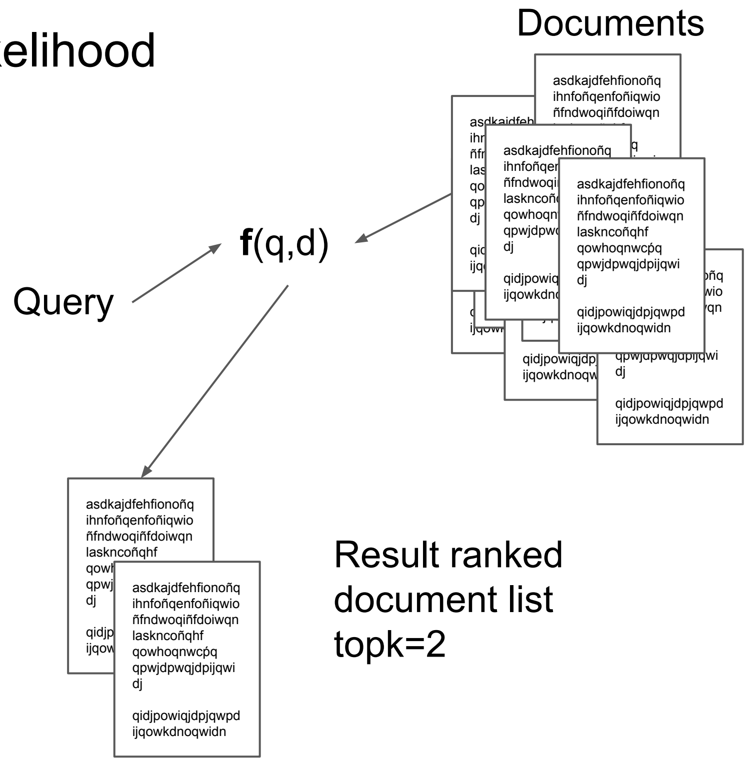
Key challenge: how to measure the likelihood that document d is relevant to query q

This is still the most used algorithm to I.R. task!!!!



I.R. models **Okapi BM25** (BM = best matching)

Key challenge: how to measure the likelihood that document d is relevant to query q



This is still the most used algorithm to I.R. task!!!!

CrossLingual I.R.???
Use Case???

Course Outline

- **Information retrieval**
  - Introduction
  - I.R. practical applications
  - I.R. for text retrieval
    - Boolean retrieval
    - Ranked retrieval
    - I.R. models (Okapi BM25)
  - **Open domain Q.A. (I.R. + Q.A.)**

This is your assignment :D
(Using Covid19 research papers)

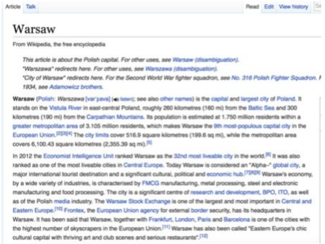
Open Domain Q.A.

Open-domain QA
SQuAD, TREC, WebQuestions, WikiMovies

Q: How many of Warsaw's inhabitants spoke Polish in 1933?

WIKIPEDIA
The Free Encyclopedia

Document Retriever


Article Talk Read Edit View history

Warsaw

From Wikipedia, the free encyclopedia

This article is about the Polish capital. For other uses, see Warsaw (disambiguation).
"Warszawa" redirects here. For other uses, see Warszawa (disambiguation).
"City of Warsaw" redirects here. For the Second World War fighter squadron, see No. 316 Polish Fighter Squadron. F
1934, see Adamowicz brothers.

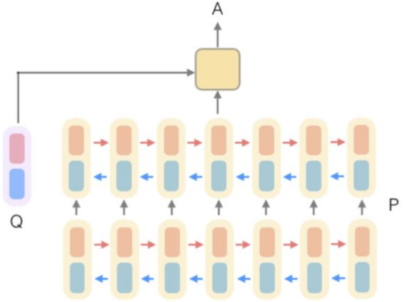
Warsaw (Polish: *Warszawa* [varˈʂava] listen); see also other names) is the capital and largest city of Poland. It stands on the Vistula River in east-central Poland, roughly 260 kilometres (160 mi) from the Baltic Sea and 300 kilometres (190 mi) from the Carpathian Mountains. Its population is estimated at 1.750 million residents within a greater metropolitan area of 3.105 million residents, which makes Warsaw the 9th most-populous capital city in the European Union.[2][3][4] The city limits cover 516.9 square kilometres (199.6 sq mi), while the metropolitan area covers 6,100.43 square kilometres (2,355.39 sq mi).[5]

In 2012 the Economist Intelligence Unit ranked Warsaw as the 32nd most liveable city in the world.[6] It was also ranked as one of the most liveable cities in Central Europe. Today Warsaw is considered an "Alpha–" global city, a major international tourist destination and a significant cultural, political and economic hub.[7][8][9] Warsaw's economy, by a wide variety of industries, is characterised by FMCG manufacturing, metal processing, steel and electronic manufacturing and food processing. The city is a significant centre of research and development, BPO, ITO, as well as of the Polish media industry. The Warsaw Stock Exchange is one of the largest and most important in Central and Eastern Europe.[10] Frontex, the European Union agency for external border security, has its headquarters in Warsaw. It has been said that Warsaw, together with Frankfurt, London, Paris and Barcelona is one of the cities with the highest number of skyscrapers in the European Union.[11] Warsaw has also been called "Eastern Europe's chic cultural capital with thriving art and club scenes and serious restaurants".[12]

Document Reader

833,500

Information Retrieval Model (I.R.)



BERT
fine-tuned for Q.A.

Thanks!

ander.barrena@ehu.eus
@4nderB

Lab Time! http://ixa2.si.ehu.eus/~jibloleo/nlpapp2

- **[8.IR_for_covid19]**
  - Edited from https://www.kaggle.com/aotegi/neural-question-answering-for-cord19-task8
    - Jon Ander Campos and Arantxa Otegi (winners of the task :D)
  - Index a set of covid related passages (from scientific papers)
  - Perform I.R. task and retrieve the **n** most relevant documents given a query
  - Test if retrieved documents are relevant
    - Persistence of virus on surfaces of different materials
    - Range of incubation periods for the disease in humans
  - **Optional:** change “domain” and use your own data, for example SQuAD passages, Wikipedia abstracts…
  - Complete the assignment using lab. code
  - Download the data in /labs/data/**passages** (176M)