# Cross-Language Information Retrieval (CLIR)

601.764

2/9/2023

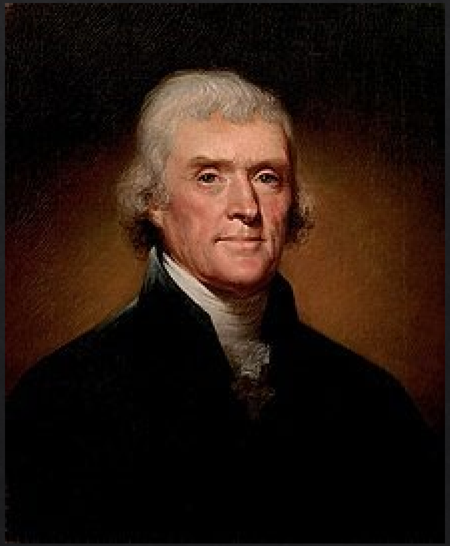
# Information Retrieval

Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).

- Manning, Raghavan, and Schütze 2008

# Library Science

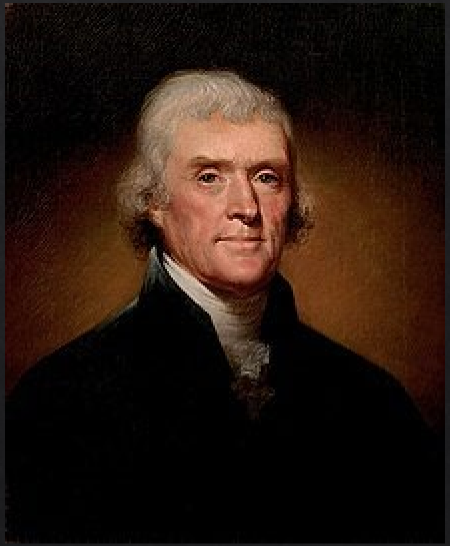
- Advice on Establishing a Library, *Gabriel Naudé* 1627
- Thomas Jefferson
  - Topics, not Alphabetical
  - Monticello → Library of Congress[1]
- Columbia University Scholar, Melvil Dewey 1887



1 Emblidge 2014

# Library Science

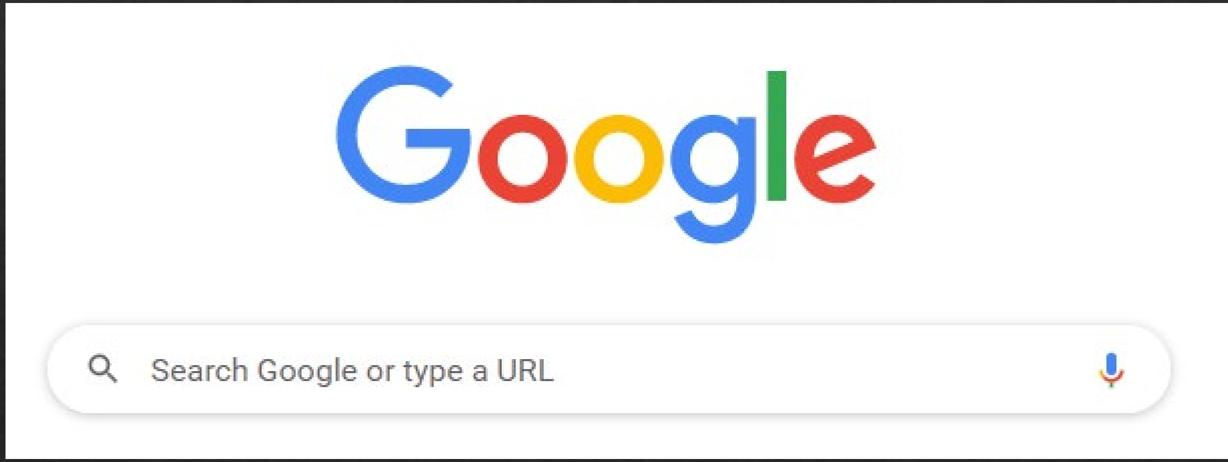
- Advice on Establishing a Library, *Gabriel Naudé* 1627
- Thomas Jefferson
  - Topics, not Alphabetical
  - Monticello → Library of Congress[1]
- Columbia University Scholar, Melvil Dewey 1887



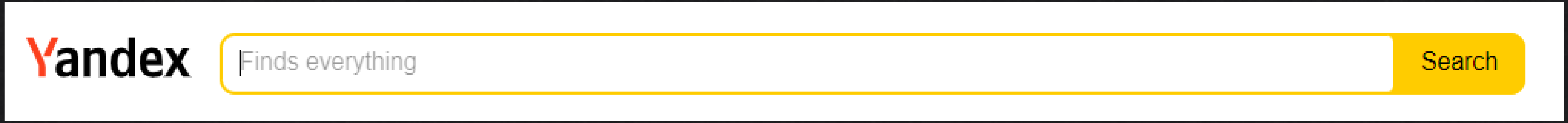
INDICES

1 Emblidge 2014

# Modern Search Engines



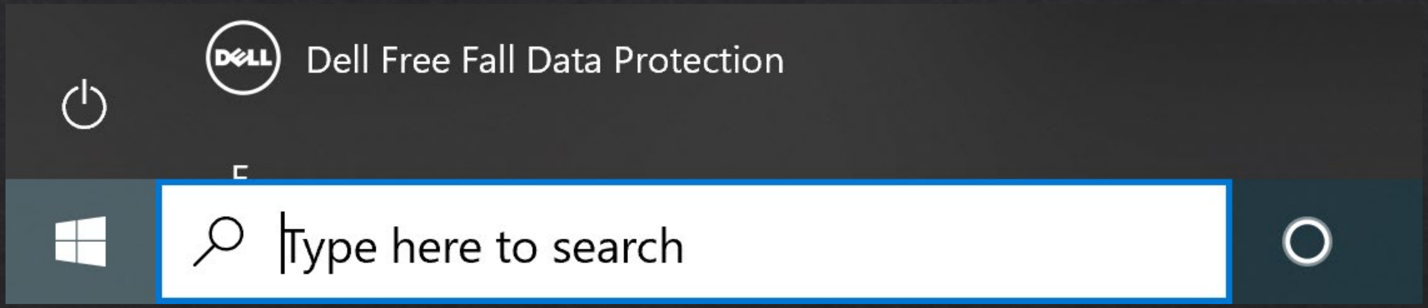


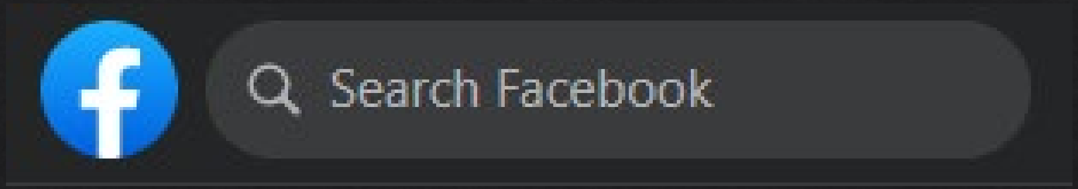



# Modern Search Engines

- Much more than just text matching
- Page Rank (Page et al., 1999)
- Click Logs

# Other IR Systems



# Information Need

- Political leaders of Russia

# Topics

- Political leaders of Russia
  - Russian Presidents
  - Soviet Premiers
  - Czars

# Queries

- Political leaders of Russia
  - Russian Presidents
  - Soviet Premiers
    - Who was the last soviet premier?
    - First soviet leader
  - Czars

# Collections

- Large number of documents
- Frequently text (ignoring image, speech, video, etc. today)
- Annotated by humans for “relevance”

# Boolean Retrieval

| | 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 | |
| ... | | | | | | | |

► **Figure 1.1** A term-document incidence matrix. Matrix element $(t, d)$ is 1 if the play in column $d$ contains the word in row $t$, and is 0 otherwise.

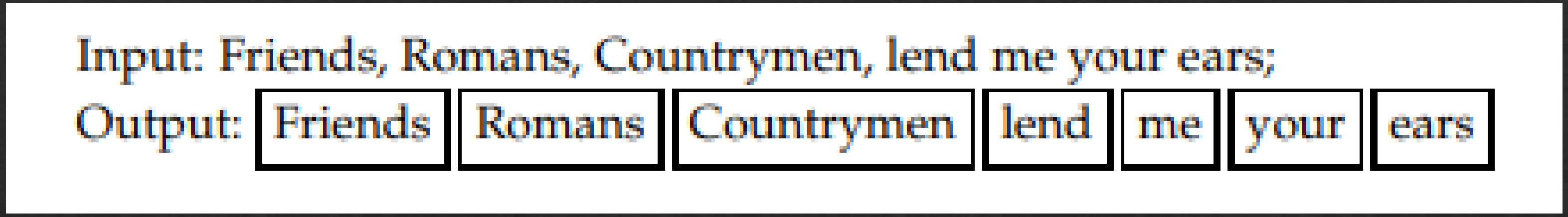
- Manning, Raghavan, and Schütze 2008

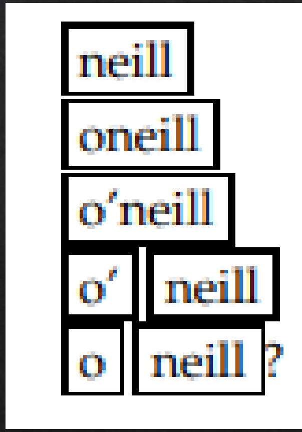
# Index Construction

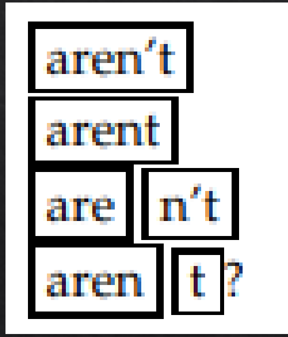
- Collect Documents
- Tokenize the Text
- Linguistic Preprocessing
- Index Documents

Manning, Raghavan, and Schütze 2008

# Tokenization




Manning, Raghavan, and Schütze 2008

# Stemming

| Rule | | | Example | | |
|---|---|---|---|---|---|
| SSES | → | SS | caresses | → | caress |
| IES | → | I | ponies | → | poni |
| SS | → | SS | caress | → | caress |
| S | → | | cats | → | cat |



Manning, Raghavan, and Schütze 2008
Porter 1980

# Index Construction

- Collect Documents
- Tokenize the Text
- Linguistic Preprocessing
- Index Documents

Language Specific!

Manning, Raghavan, and Schütze 2008

# Tokenization

莎拉波娃现在居住在美国东南部的佛罗里达。今年 4 月
9 日，莎拉波娃在美国第一大城市纽约度过了 1 8 岁生
日。生日派对上，莎拉波娃露出了甜美的微笑。

和尚

► **Figure 2.4** Ambiguities in Chinese word segmentation. The two characters can be treated as one word meaning 'monk' or as a sequence of two words meaning 'and' and 'still'.

Manning, Raghavan, and Schütze 2008

# Morphology

El Gato

La Gata

# Cross-Language Information Retrieval
# Term-Document Index

| | Declaration of Independence | US Constitution | Bill of Rights | Gettysburg Address |
|---|---|---|---|---|
| Il | | | | |
| y | | | | |
| a | | | | |
| quatre | | | | |
| vingt | | | | |
| et | | | | |
| sept | | | | |
| ans | | | | |

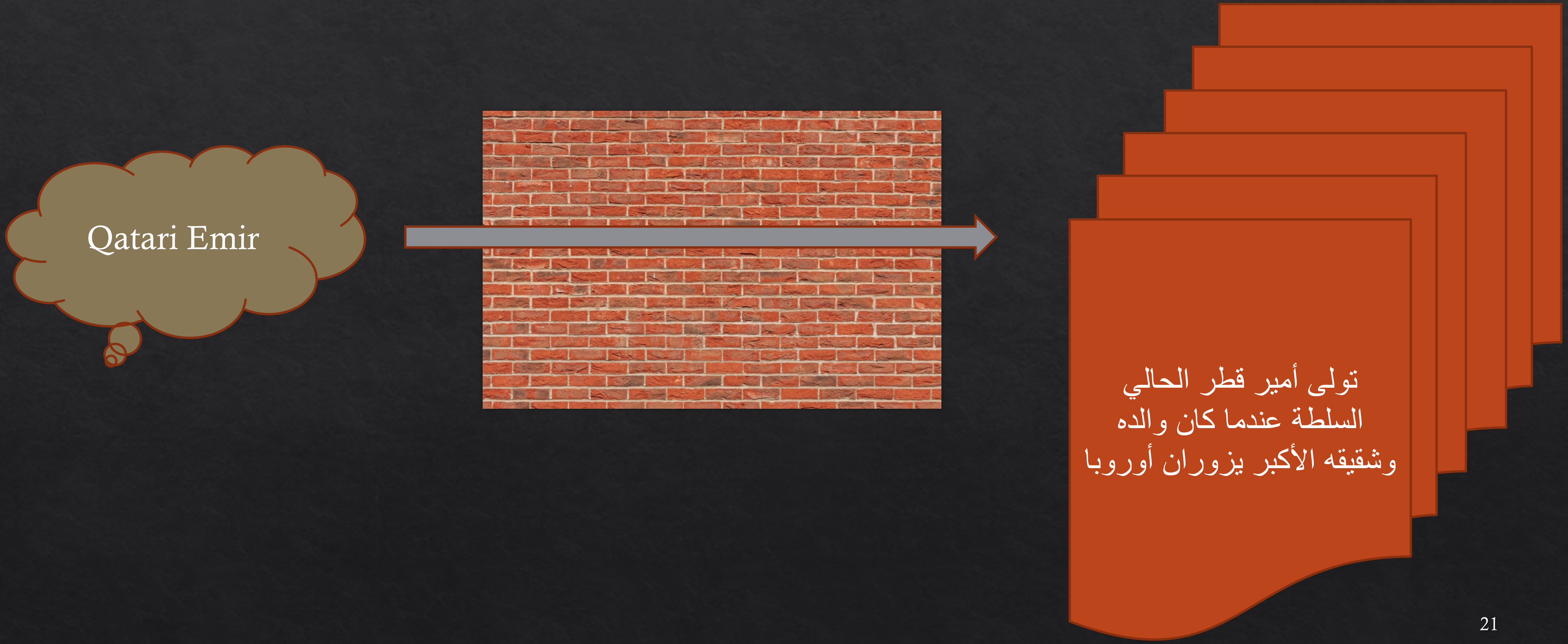
# Cross-Language Information Retrieval
# Term-Document Index

| | **Declaration of Independence** | **US Constitution** | **Bill of Rights** | Gettysburg Address |
|---|---|---|---|---|
| Il | | | | |
| y | | | | |
| a | | | | |
| quatre | | | | |
| vingt | | | | |
| et | | | | |
| sept | | | | |
| ans | | | | |

Stop word?

Month?

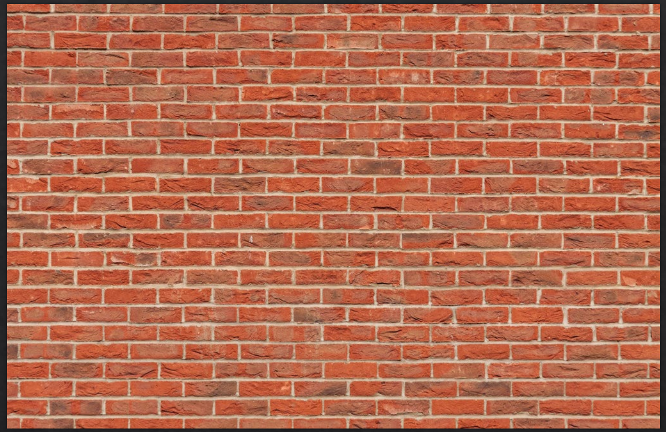
# Crossing the Language Barrier



# Crossing the Language Barrier

Qatari Emir

أمير قطر



تولى أمير قطر الحالي
السلطة عندما كان والده
وشقيقه الأكبر يزوران أوروبا

# Crossing the Language Barrier

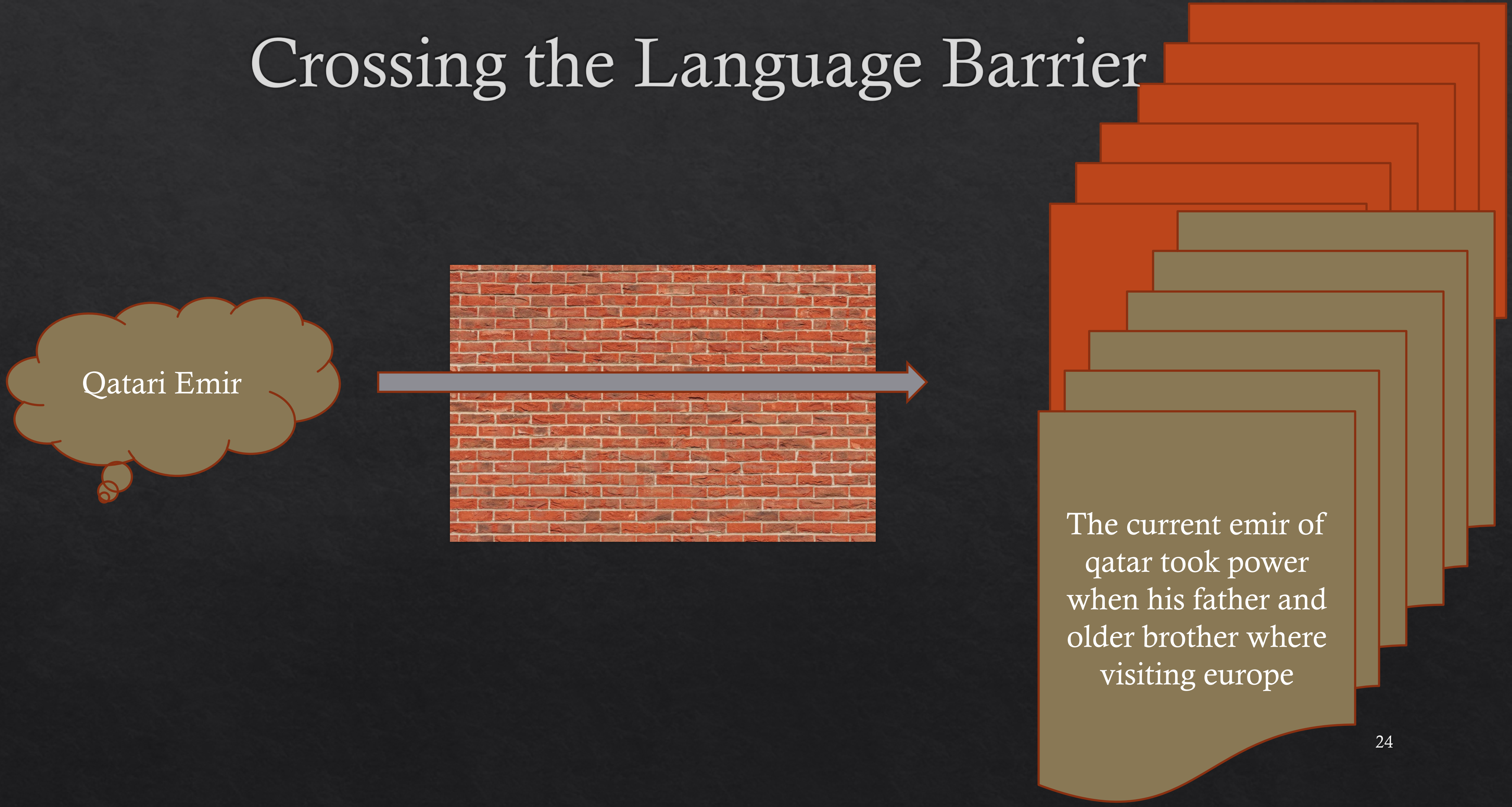
Qatari Emir

أمير قطر

The current emir of qatar took power when his father and older brother where visiting europe

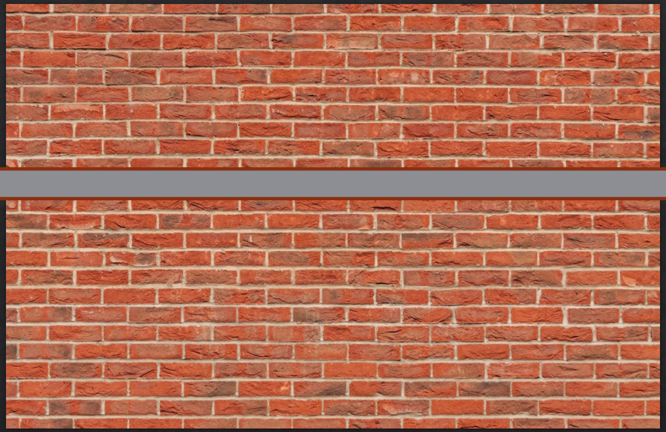
تولى أمير قطر الحالي
السلطة عندما كان والده
وشقيقه الأكبر يزوران أوروبا

# Crossing the Language Barrier



# Crossing the Language Barrier

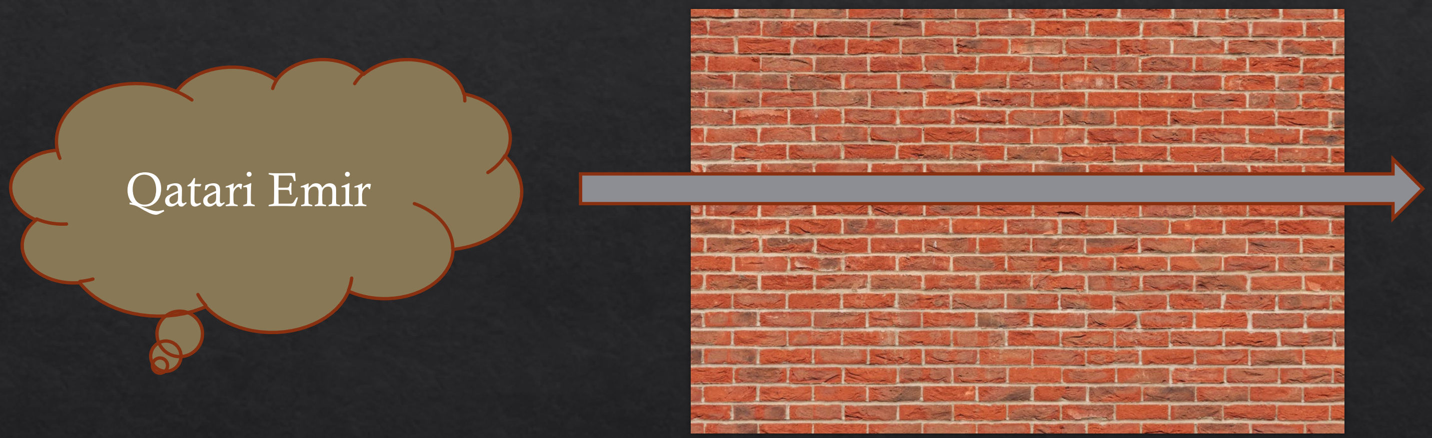
Qatari Emir



The current emir of qatar took power when his father and older brother where visiting europe

**Slower (All of your Corpus)**

# Crossing the Language Barrier




The current emir of qatar took power when his father and older brother where visiting europe

**Slower (All of your Corpus) Works better***

## Multilingual Information Retrieval

# Key Question: How do you consistently rank across languages?

The current emir of
qatar took power
when his father and
older brother where
visiting europe

تولى أمير قطر الحالي
السلطة عندما كان والده
وشقيقه الأكبر يزوران أوروبا

卡塔爾現任埃米爾在
他的父親和哥哥訪問
歐洲時掌權

कतारस्य वर्तमानः
अमीरः सत्तां गृहीतवान्
यदा तस्य पिता
अग्रजः च यत्र
यूरोपदेशं गच्छति स्म

# Multilingual Information Retrieval

## **Are these just translations?**

The current emir of
qatar took power
when his father and
older brother where
visiting europe

تولى أمير قطر الحالي
السلطة عندما كان والده
وشقيقه الأكبر يزوران أوروبا

卡塔爾現任埃米爾在
他的父親和哥哥訪問
歐洲時掌權

कतारस्य वर्तमानः
अमीरः सत्तां गृहीतवान्
यदा तस्य पिता
अग्रजः च यत्र
यूरोपदेशं गच्छति स्म

# Multilingual Information Retrieval

## **Next week's lecture…**

The current emir of
qatar took power
when his father and
older brother where
visiting europe

تولى أمير قطر الحالي
السلطة عندما كان والده
وشقيقه الأكبر يزوران أوروبا

卡塔爾現任埃米爾在
他的父親和哥哥訪問
歐洲時掌權

कतारस्य वर्तमानः
अमीरः सत्तां गृहीतवान्
यदा तस्य पिता
अग्रजः च यत्र
यूरोपदेशं गच्छति स्म

# Sparse Retrieval

| | 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 | |
| ... | | | | | | | |

► **Figure 1.1** A term-document incidence matrix. Matrix element $(t, d)$ is 1 if the play in column $d$ contains the word in row $t$, and is 0 otherwise.

| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | | | | | | | | | | | | | | | | | |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | | | | | | | | | | | | | | | | | |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | | | | | | | | | | | | | | | | | |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | | | | | | | | | | | | | | | | | |
| 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | | | | | | | | | | | | | | | | | |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | | | | | | | | | | | | | | | | | |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | | | | | | | | | | | | | | | | | |
| 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | | | | | | | | | | | | | | | | | |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | | | | | | | | | | | | | | | | | |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | | | | | | | | | | | | | | | | | |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | | | | | | | | | | | | | | | | | |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | | | | | | | | | | | | | | | | | |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | | | | | | | | | | | | | | | | | |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | | | | | | | | | | | | | | | | | |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | | | | | | | | | | | | | | | | | |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

@ChrisVWarren

# Language Agnostic (Given Tokenization,…)

## BM 25

# Tldr; Score Terms by How Often they appear in a Document vs. Entire Corpus

- Okapi Best Match 25
- Robertson et al., 1995

\# times $q_i$ appears in D

Query

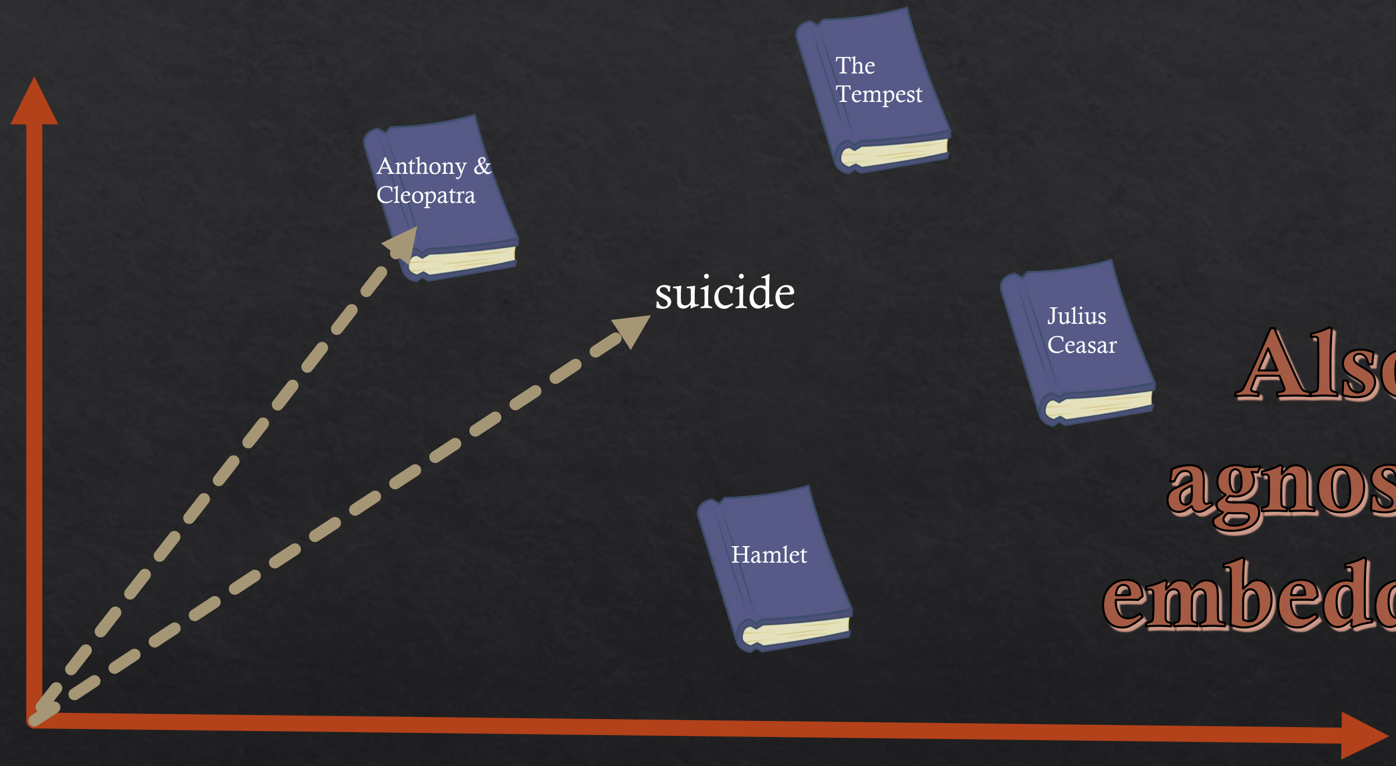
$$Score(D, Q) = \sum_{i=1}^{n} IDF(qi)\left( \frac{f(qi, D)(k_1 + 1)}{f(q_i, D) + k_1(1 - b + b\frac{|D|}{avg|D|})} \right)$$

Document

Score for how often $q_i$ appears in the corpus

Manually Tuned Hyperparameters

# Dense Retrieval

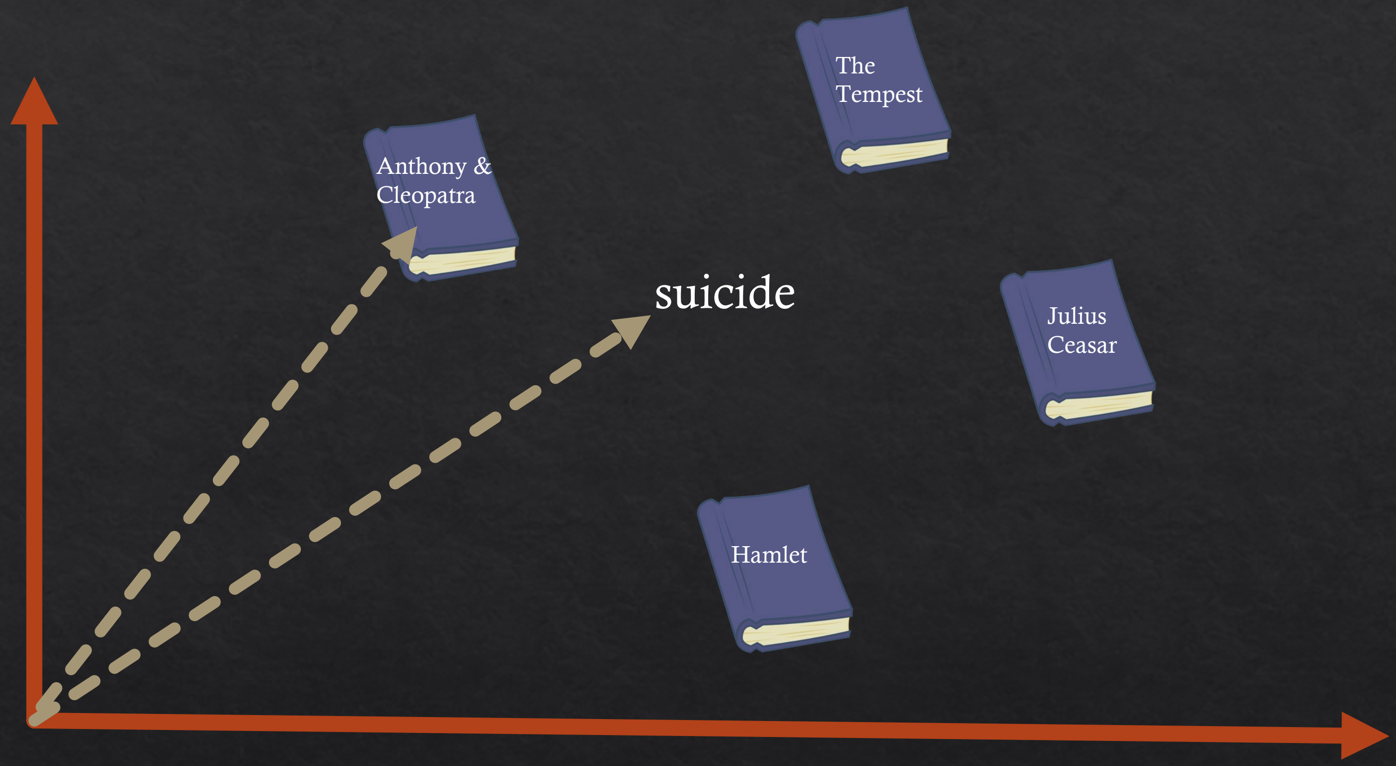



**Slow ☹**

**Also, language agnostic assuming embedding algorithm**

Continuous Vector Space

# Vector Space




*TF.IDF* is also a vector space model!

# Independence Assumption

## Vector Space

- Boolean Document Model
- *TF* Document Model
- *TF.IDF* Document Model

```
Doc1: It is a sunny day in Karlsruhe.
Doc2: It rains and rains and rains the whole day.
```

| | *Boolean* | | *TF* | | *TF.IDF* | |
|---|---|---|---|---|---|---|
| *Term* | Doc1 | Doc2 | Doc1 | Doc2 | Doc1 | Doc2 |
| sunny | 1 | 0 | 1 | 0 | $1 \log 2/1 = 0.7$ | 0.0 |
| day | 1 | 1 | 1 | 1 | $1 \log 2/2 = 0.0$ | $1 \log 2/2 = 0.0$ |
| Karlsruhe | 1 | 0 | 1 | 0 | $1 \log 2/1 = 0.7$ | 0.0 |
| rains | 0 | 1 | 0 | 3 | 0.0 | $3 \log 2/1 = 2.1$ |

Sorg 2011

# Independence Assumption

- Sparse Retrieval frequently assumes that terms occur independently in document
  - Simplifying, but worked well
  - Pre-Compute
- Dense Retrieval: Contextual Language Models

# Reranking


Initial Sparse Retrieval

Declaration of Independence
Hamlet
Othello
Constitution
Romeo & Juliet
……
Macbeth

~1,000 Documents

Dense Retrieval

Hamlet
Othello
Romeo & Juliet
Macbeth
……
Declaration of Independence
Constitution

# Encode Using a Pretrained Language Model

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **0.07** | **3.24** | **0.30** | **-1.50** | **0.77** | **0.82** | **0.24** | **1.40** |

**Slow.**
**Need to compute query & document at inference time**

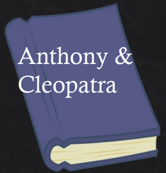
Suicide | | | Two households, both alike in dignity. In fair Verona ….

# Twin Towers

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 0.00 | -0.35 | 7.11 | 2.50 | 0.89 | 2.24 | 0.05 | 1.08 |
| 0.01 | 0.24 | 1.30 | 2.50 | 0.13 | -0.32 | 0.24 | 1.40 |

0.642

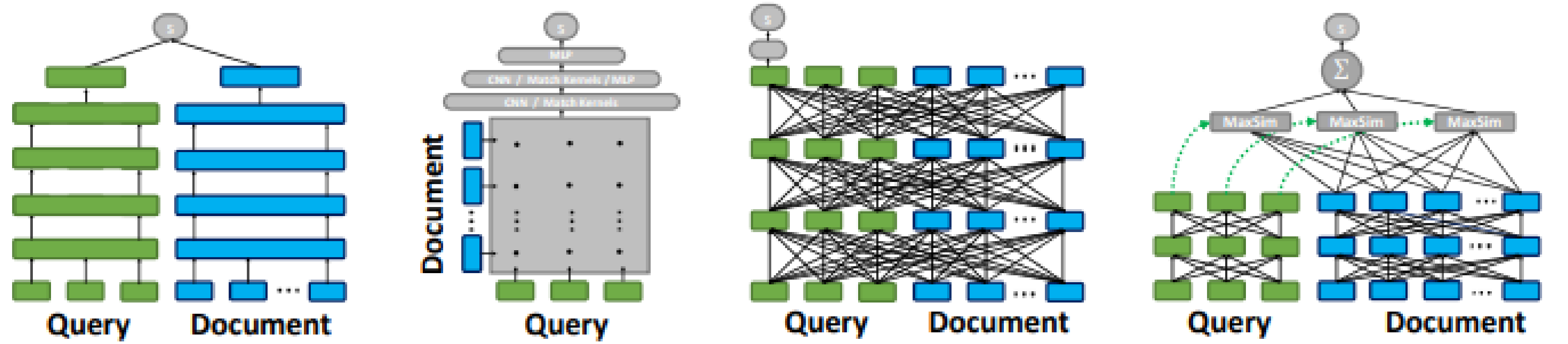
Can still be slow depending on similarity function




suicide

# ColBERT

- Khattab and Zaharia 2020
- Twin Tower Dense Retrieval
- MaxSim Operator (Independent Gradients for Encoders)

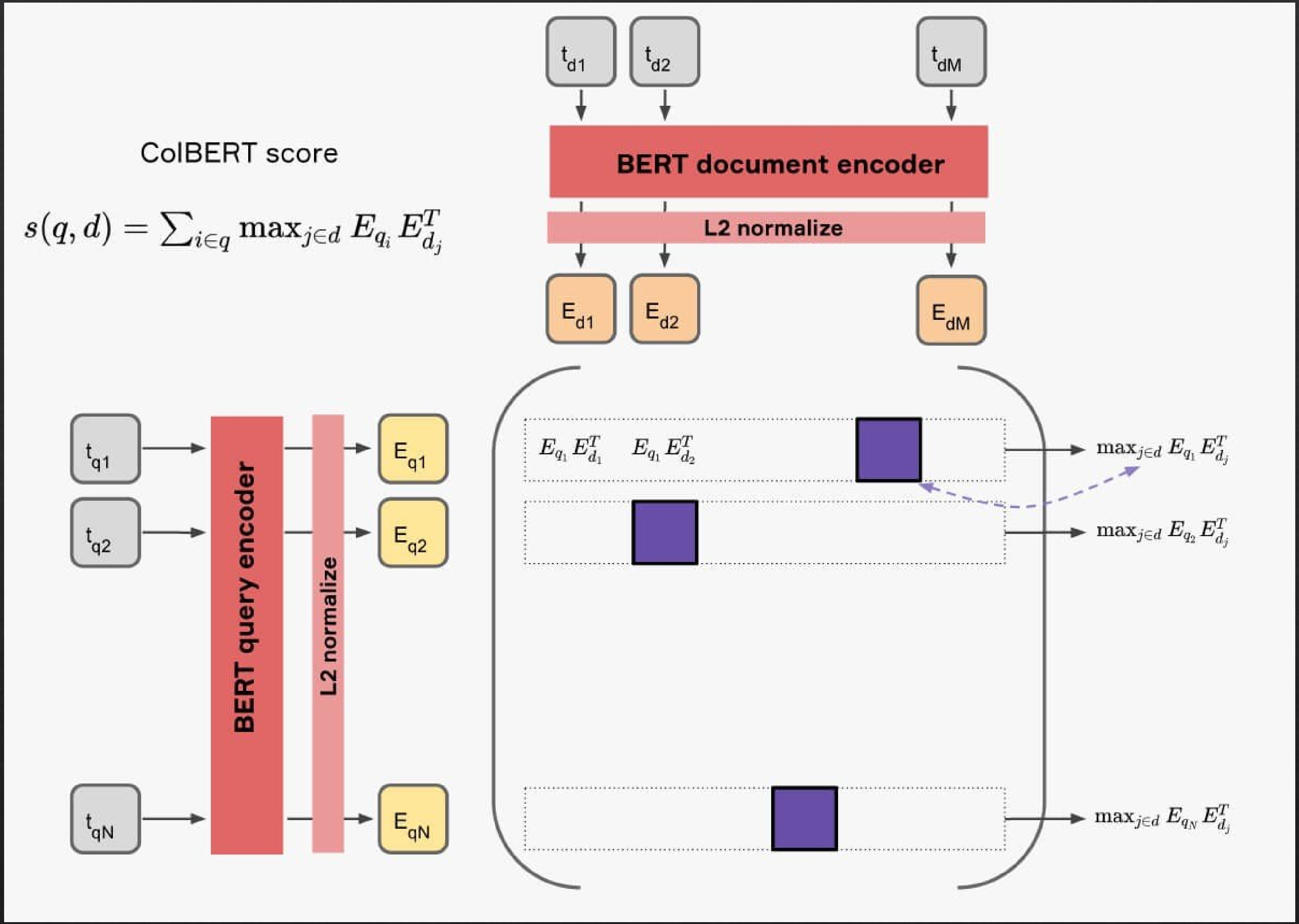



(a) Representation-based Similarity (e.g., DSSM, SNRM)

(b) Query-Document Interaction (e.g., DRMM, KNRM, Conv-KNRM)

(c) All-to-all Interaction (e.g., BERT)

(d) Late Interaction (i.e., the proposed ColBERT)




https://medium.com/@varun030403/colbert-a-complete-guide-1552468335ae

# NeuCLIR



Devlin et al., 2018

# mBERT (Multilingual BERT)

NeuCLIR

Devlin et al., 2018

# ColBERT-X (Nair et al, 2022)

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 0.00 | -0.35 | 7.11 | 2.50 | 0.89 | 2.24 | 0.05 | 1.08 |

0.642

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 0.01 | 0.24 | 1.30 | 2.50 | 0.13 | -0.32 | 0.24 | 1.40 |

**Translate Queries**

**Translate Documents**

**Neither ☺**


انتحار

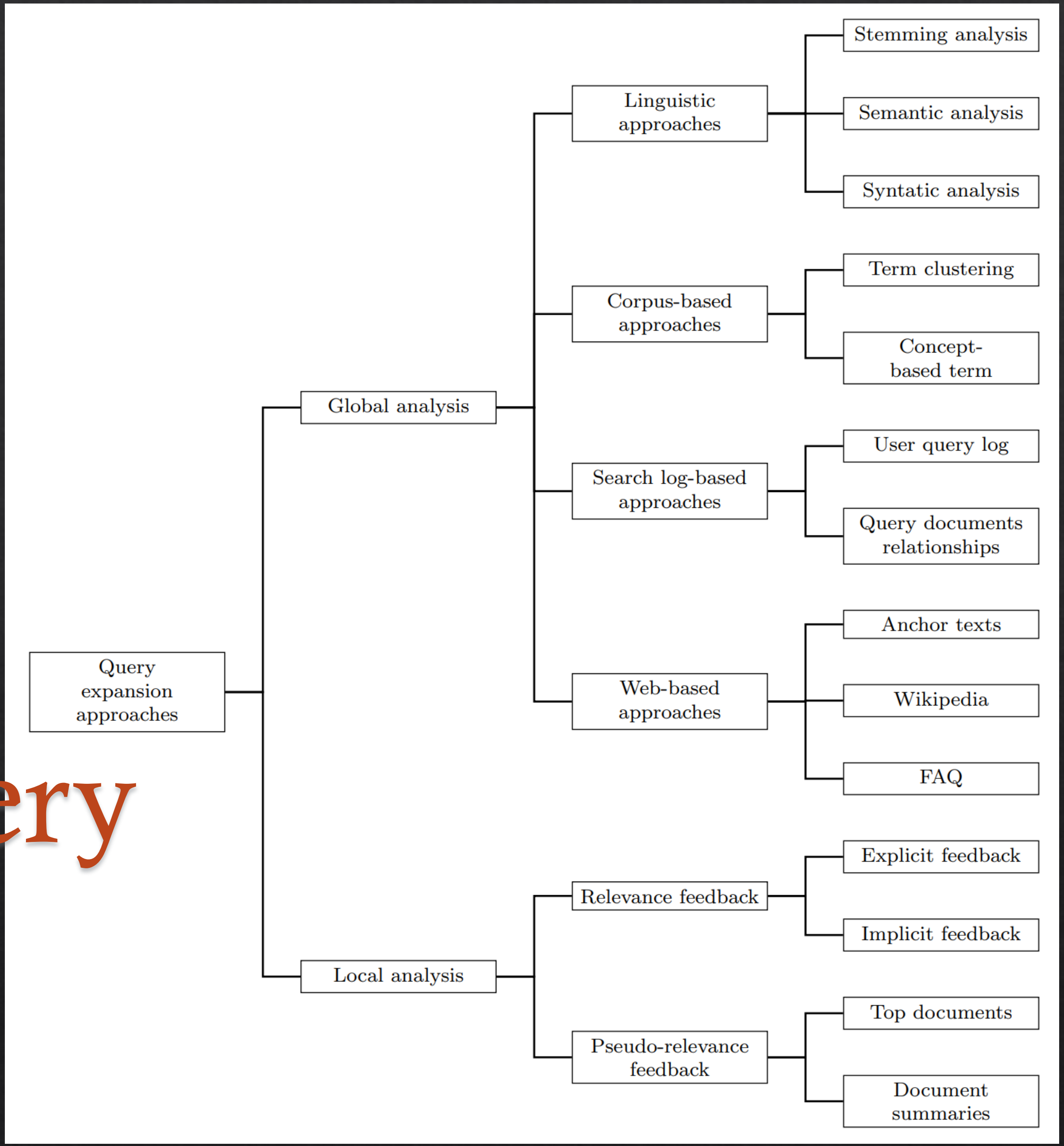
# Query Expansion

- Additional Terms
- Frequently given less weight

# Query Expansion

- Background Knowledge
- Relevance Feedback
- Sorg and Cimiano (Chapter 11)

Translation of Query




Azad and Deepak 2019

# Query Expansion

## How do we make this multilingual?

Azad and Deepak 2019

| Type of Data Sources | Data Sources | Term Extraction Methodology |
|---|---|---|
| Documents Used in Retrieval Process | Clustered terms | Clustering of terms and documents from sets of similar objects |
| | Corpus or Collection based data sources | Terms collection from specific domain knowledge |
| Hand Built Knowledge Resources | WordNet & Thesaurus | Word sense and synset |
| | ConceptNet & Knowledge bases | Common sense knowledge and Freebase |
| | Wikipedia or DBpedia | Articles, titles & hyper links |
| External Text Collections and Resources | Anchor texts | Adjacent terms in anchor text or text extraction from anchor tags |
| | Query logs or User logs | Historical records of user queries registered in the query logs of search engine |
| | External corpus | Nearby terms in word embedding framework |
| Hybrid Data Sources | Top-ranked documents & multiple sources | All terms in top retrieved documents |

# Summary

- Many IR methods rely on language specific parts of a pipeline
- Numerous linguistic challenges exist for CLIR
- Neural Networks have opened up new possibilities
- Active area of research
- Not enough collections