

67-300 SEARCH ENGINES

INTRODUCTION

LECTURER: JOAO PALOTTI (JPALOTTI@ANDREW.CMU.EDU)

13TH MARCH 2017

CLASS SUMMARY

- ▶ Course rules
- ▶ Course introduction
- ▶ Introduction to Information Retrieval
- ▶ Practice Exercises

HOW IS THIS COURSE GOING TO BE?

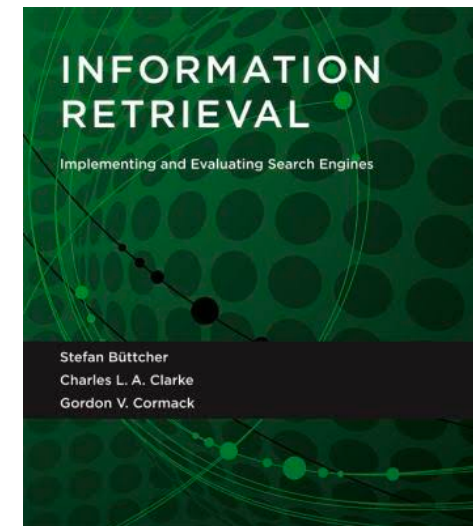
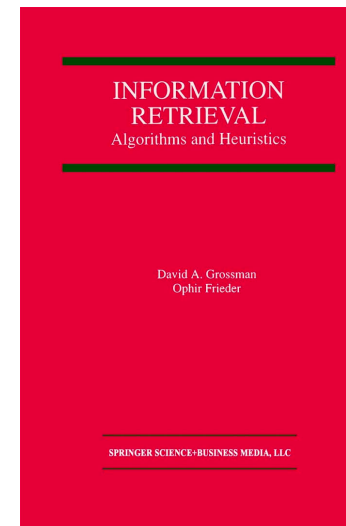
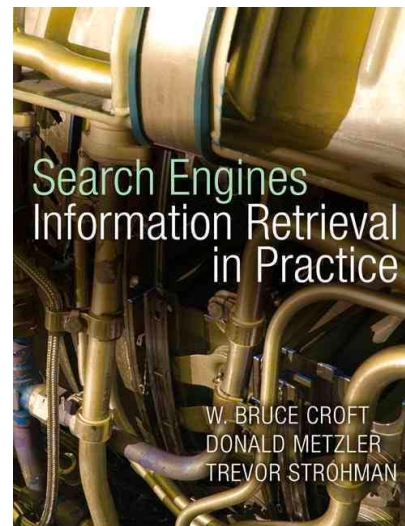
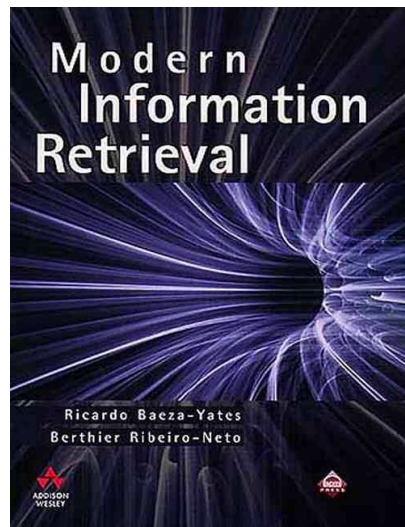
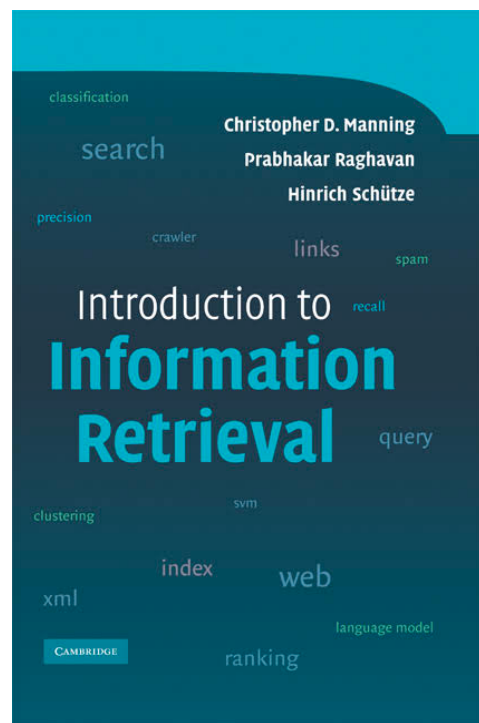
- ▶ Sit, relax and **ask questions!!!!**
- ▶ **You participation is encouraged and expected!**
- ▶ In class exercises whenever possible
- ▶ Quizzes at the end of the class to review main concepts (piazza)

PIAZZA

- ▶ We will use Piazza for:
 - ▶ Communication
 - ▶ Submit your weekly quiz
 - ▶ Submit your project deliverables

TEXT BOOK

- ▶ There are many excellent books in this area:



→ <http://nlp.stanford.edu/IR-book/>

POLICIES

- ▶ Submit homework via Piazza before class starts
- ▶ Do not wait till is too late to start your project
 - ▶ Every day that you are late you will lose 10% of the maximum score.
- ▶ Do not copy code from any other student
- ▶ Do not copy reports from other students

WHAT IS THIS MINI COURSE GOING TO BE ABOUT?

- ▶ Understand what is behind any search bar:
 - ▶ Overview of many and many years of empirical research on text processing

Google

SEZNAM.CZ

شامرا SHAMRA.SY

bing

Yandex

YAHOO!

WHAT IS THIS MINI COURSE GOING TO BE ABOUT?

- ▶ Understand what is behind any search bar:
 - ▶ Overview of many and many years of empirical research on text processing

Google

SEZNAM.CZ

شامرا SHAMRA.SY

bing

Yandex

amazon.com

YAHOO!

PubMed

Alibaba.com

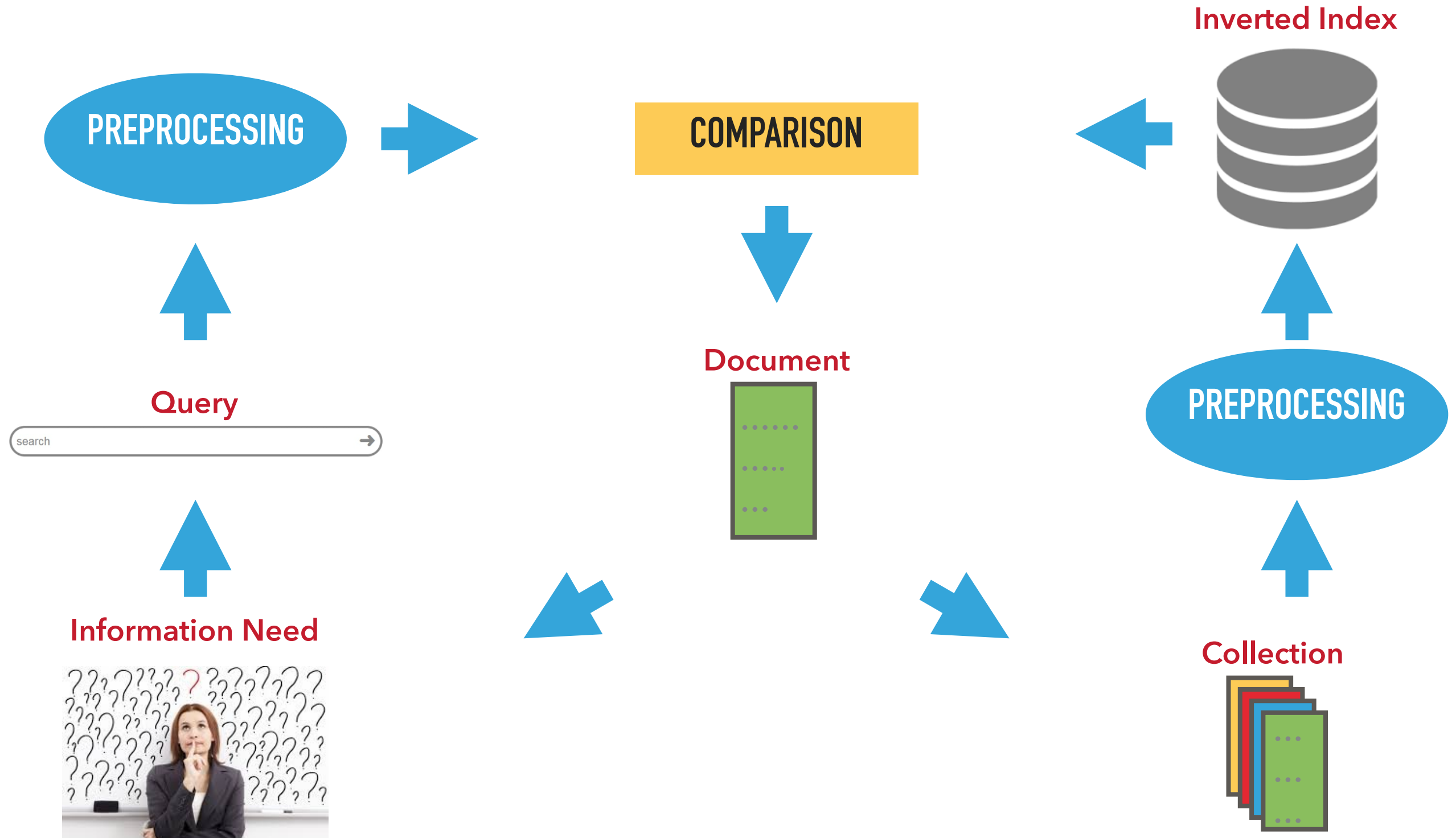
WHAT IS THIS MINI COURSE GOING TO BE ABOUT?

- ▶ Understand what is behind any search bar:
 - ▶ Overview of many and many years of empirical research on text processing
- ▶ Be able to process text using Python:
 - ▶ Quick prototyping, excellent for text processing
 - ▶ Practice programming

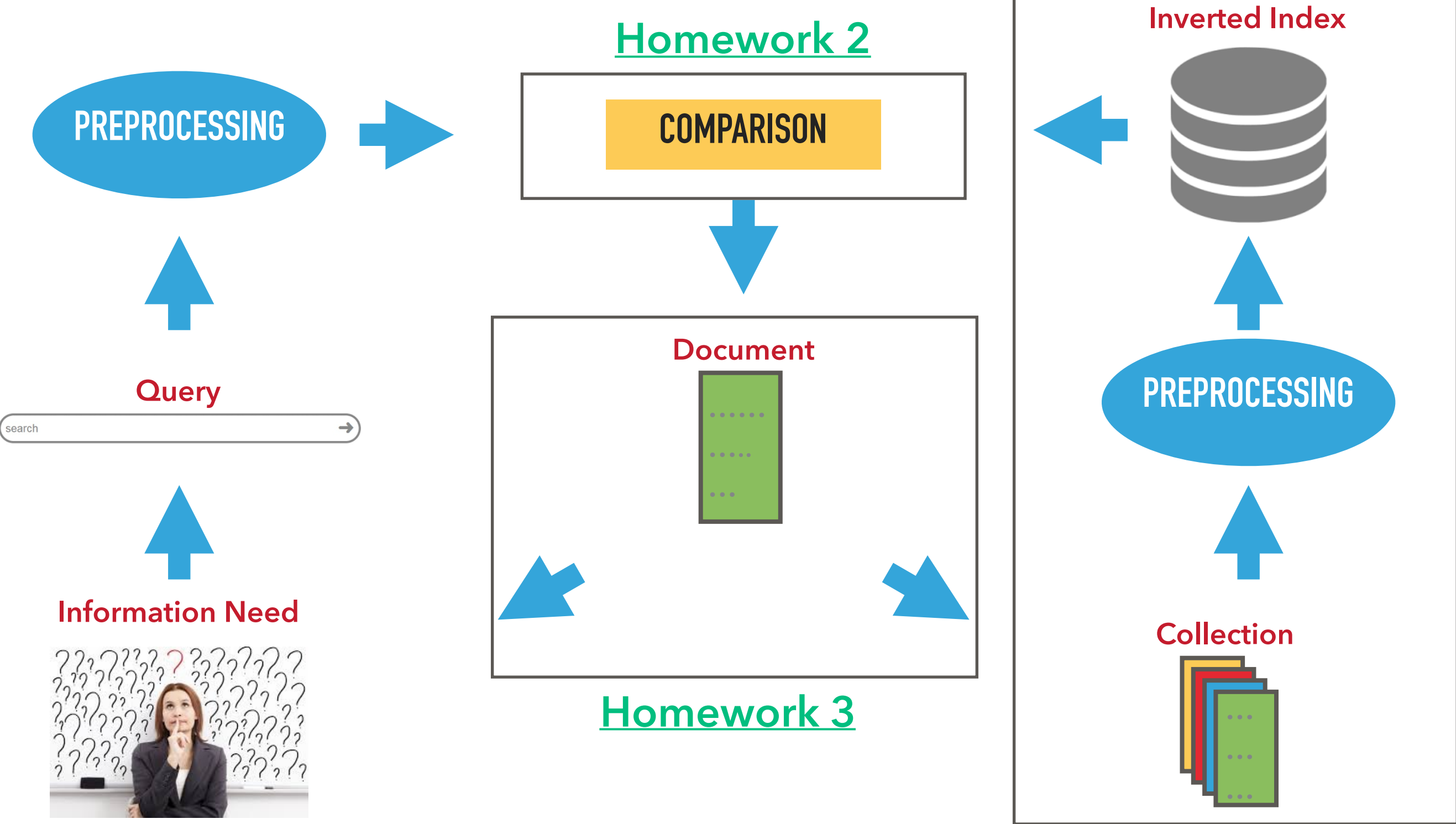
GRADING

- ▶ 3 small projects (60)
 - ▶ First (15): Text preprocessing Index creation
 - ▶ Second (25): Search models
 - ▶ Third (20): Evaluation
- ▶ Class Participation (10)
- ▶ Final exam (30)

BIG PICTURE

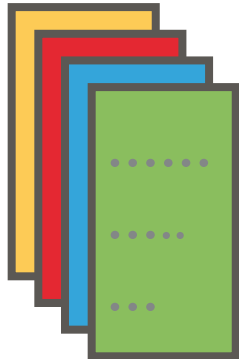


BIG PICTURE

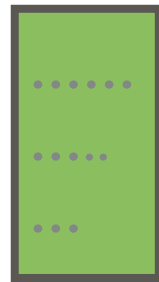


LET'S START

Collection of Documents



Document



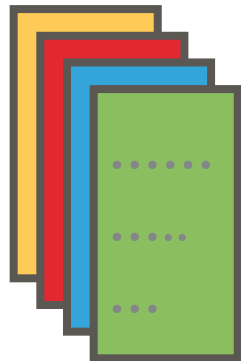
Where is Qatar?

Qatar is a country located in Western Asia, occupying the small Qatar Peninsula on the northeastern coast of the Arabian Peninsula. Its sole land border is with Saudi Arabia to the south, with the rest of its territory surrounded by the Persian Gulf.

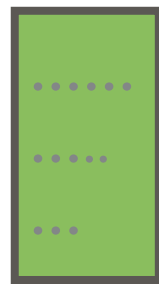
WIKIPEDIA ENTRY FOR QATAR

LET'S START

Collection of Documents



Document



Where is Qatar?

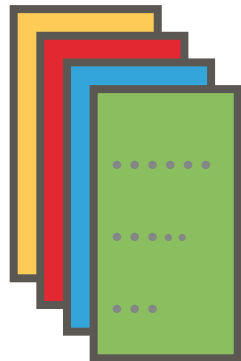
Qatar is a country located in Western Asia, occupying the small Qatar Peninsula on the northeastern coast of the Arabian Peninsula. Its sole land border is with Saudi Arabia to the south, with the rest of its territory surrounded by the Persian Gulf.

WIKIPEDIA ENTRY FOR QATAR

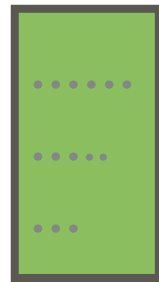
What is this text about? How do you know that?

LET'S START

Collection of Documents



Document



Where is Qatar?

Qatar is a country located in Western Asia, occupying the small Qatar Peninsula on the northeastern coast of the Arabian Peninsula. Its sole land border is with Saudi Arabia to the south, with the rest of its territory surrounded by the Persian Gulf.

WIKIPEDIA ENTRY FOR QATAR

What is this text about? How do you know that?

What if you have a query like that?

qatar location



TOKENIZATION

Where is Qatar?

Qatar is a country located in Western Asia, occupying the small Qatar Peninsula on the northeastern coast of the Arabian Peninsula. Its sole land border is with Saudi Arabia to the south, with the rest of its territory surrounded by the Persian Gulf.

WIKIPEDIA ENTRY FOR QATAR

Original Text

TOKENIZATION

Where is Qatar?

Qatar is a country located in Western Asia, occupying the small Qatar Peninsula on the northeastern coast of the Arabian Peninsula. Its sole land border is with Saudi Arabia to the south, with the rest of its territory surrounded by the Persian Gulf.

WIKIPEDIA ENTRY FOR QATAR

Original Text

1. Lower/Upper case? Always work? Think of acronyms.
2. Is punctuation important?
3. Is the order of the words important?

TOKENIZATION

Where is Qatar?

Qatar is a country located in Western Asia, occupying the small Qatar Peninsula on the northeastern coast of the Arabian Peninsula. Its sole land border is with Saudi Arabia to the south, with the rest of its territory surrounded by the Persian Gulf.

WIKIPEDIA ENTRY FOR QATAR

Original Text

BAG OF WORDS

is peninsula of
qatar south
country the peninsula
border arabia is where
qatar persian qatar
arabian western
territory gulf saudi
as the arab coast
occupying land asia
a small located is

Computer representation

1. Lower/Upper case? Always work? Think of acronyms.
2. Is punctuation important?
3. Is the order of the words important?

TOKENIZATION

Where is Qatar?

Qatar is a country located in Western Asia, occupying the small Qatar Peninsula on the northeastern coast of the Arabian Peninsula. Its sole land border is with Saudi Arabia to the south, with the rest of its territory surrounded by the Persian Gulf.

WIKIPEDIA ENTRY FOR QATAR

Original Text

BAG OF WORDS

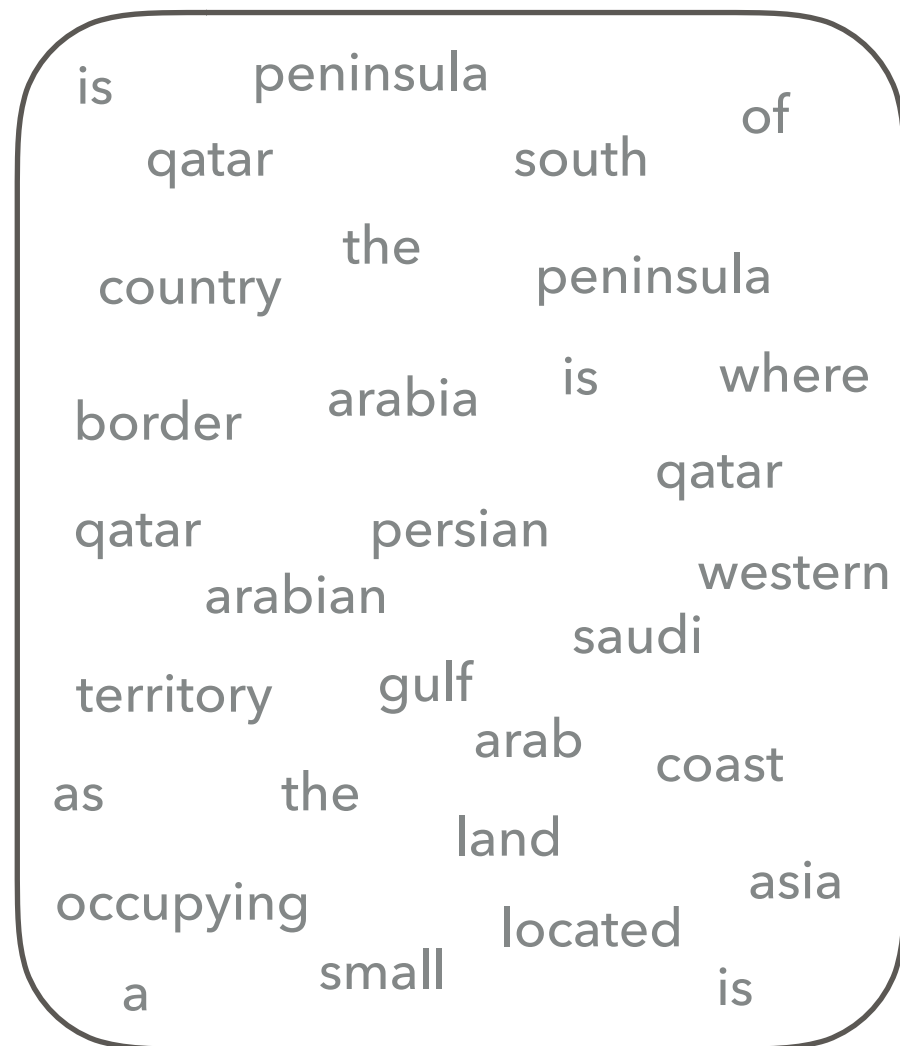
is peninsula of
qatar south
country the peninsula
border arabia is where
qatar persian qatar
arabian western
territory gulf saudi
as the arab coast
occupying land asia
a small located is

Computer representation

Does it work?

TOKENIZATION

BAG OF WORDS



Computer representation

TOKENS

Word Occurrences

31 tokens

vs

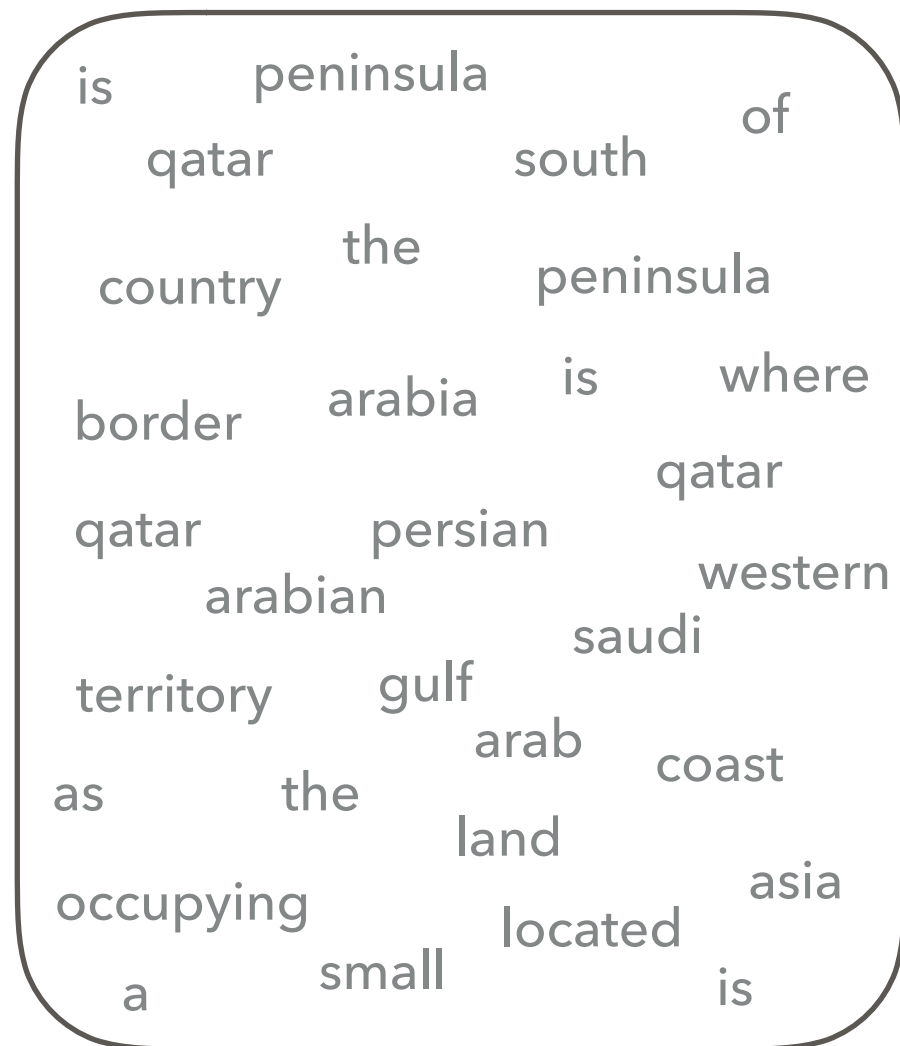
TYPES

Unique Words

26 types

TOKENIZATION

BAG OF WORDS



Computer representation

TOKENS

Word Occurrences

31 tokens

VS

TYPES

Unique Words

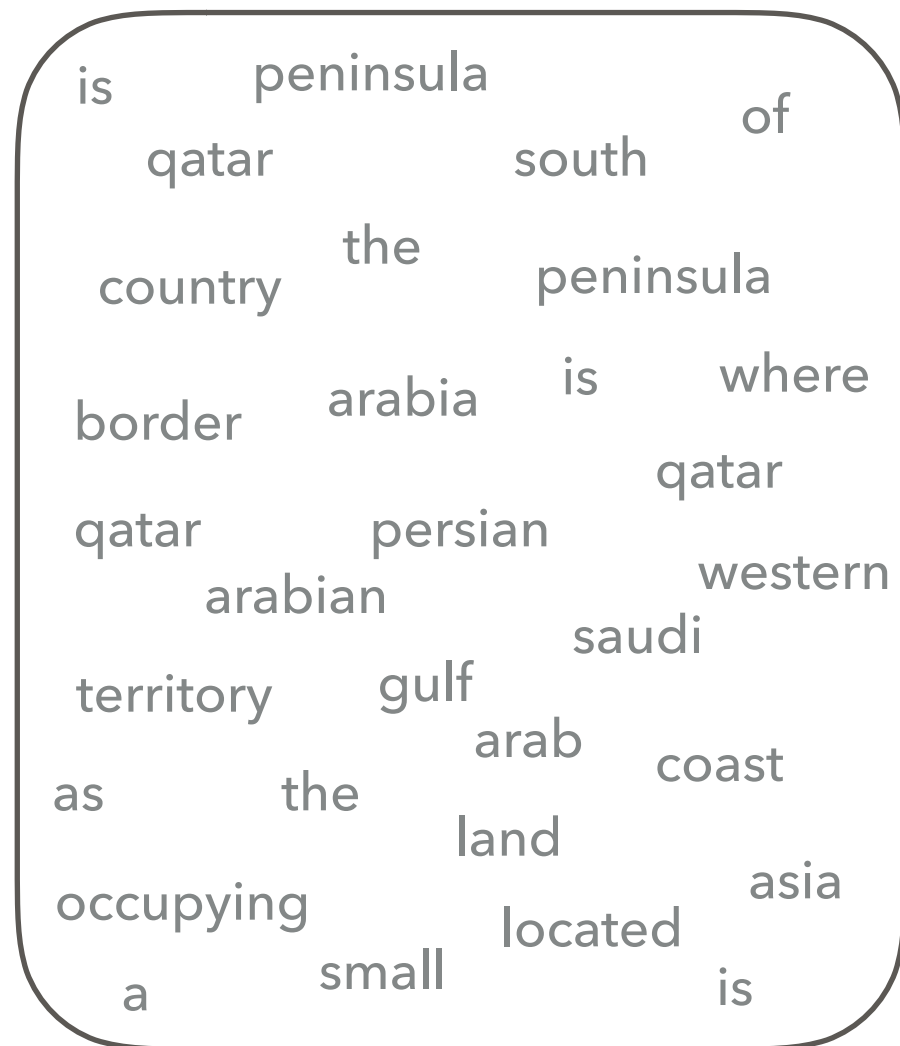
26 types

Are all words necessary?

Are all words equally important?

TOKENIZATION

BAG OF WORDS



Computer representation

TOKENS

Word Occurrences

31 tokens

vs

TYPES

Unique Words

26 types

Are all words necessary?

Are all words equally important?

No! Stopwords are usually

not important....

TOKENIZATION

BAG OF WORDS



Computer representation

TOKENS

Word Occurrences

31 tokens

vs

TYPES

Unique Words

26 types

Are all words necessary?

Are all words equally important?

No! Stopwords are usually

not important....

TOKENIZATION

BAG OF WORDS



Computer representation

TOKENS

Word Occurrences

31 tokens

vs

TYPES

Unique Words

26 types

Are all words necessary?

Are all words equally important?

No! Stopwords are usually

not important....

TOKENIZATION

BAG OF WORDS



Computer representation

What else can we do?

TOKENIZATION

BAG OF WORDS



peninsula
qatar south
country peninsula
border arabia where
qatar persian qatar
arabian western
territory gulf saudi
arab coast
land
occupying asia
small located

Computer representation

What else can we do?

arab, arabia, arabian, arabic

TOKENIZATION

BAG OF WORDS



Computer representation

What else can we do?

arab, arabia, arabian, arabic

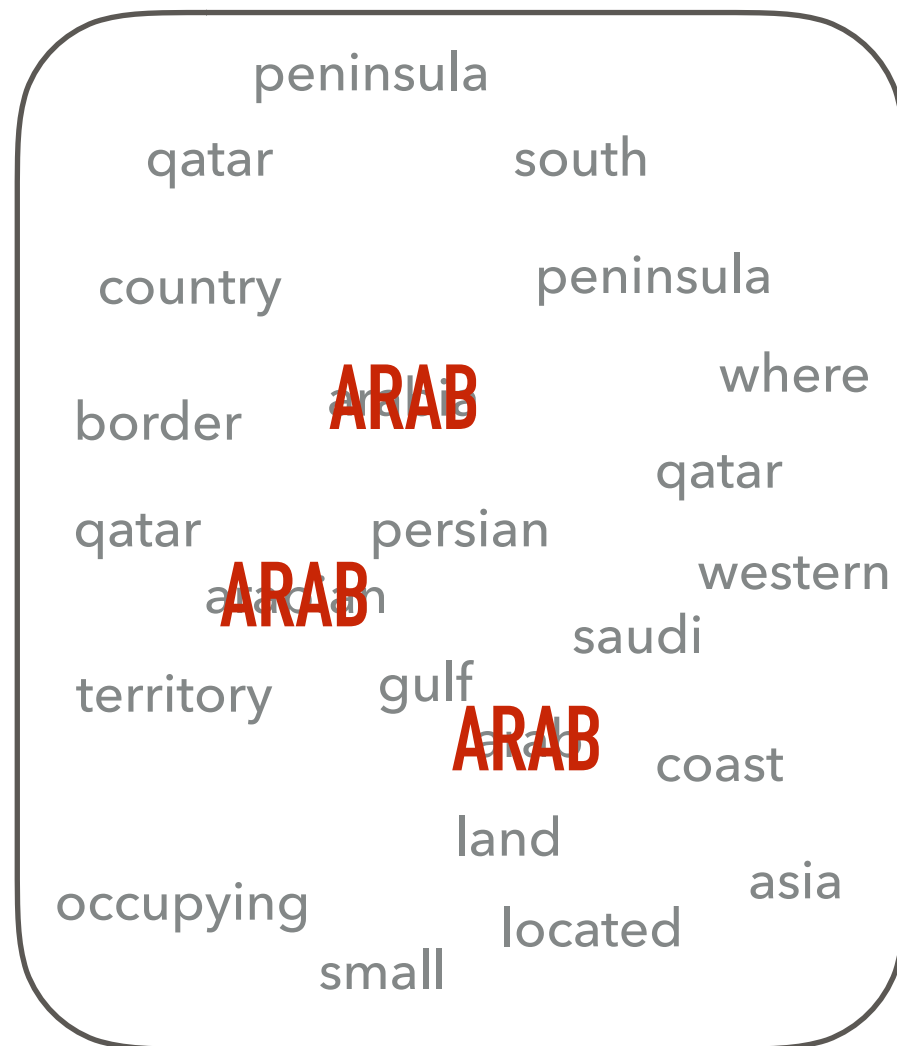


STEMMER

ARAB

TOKENIZATION

BAG OF WORDS



Computer representation

What else can we do?

arab, arabia, arabian, arabic



STEMMER

ARAB

TOKENIZATION

BAG OF WORDS



Computer representation

What else can we do?

arab, arabia, arabian, arabic



STEMMER

ARAB

Note the use of stemmer in all tokens of our example

PREPROCESSING/NORMALIZATION STEPS

► Tokenization summary:

1. Lower/Upper case
2. Split text into tokens
3. Deal with punctuation
4. Stopwords removal
5. Stemming

BAG OF WORDS (BOW)



Computer representation

Bonus: accents/diacritics; date format; British/American spelling; Equivalence classes

Language challenges (Chinese has no white space between words)

IS THERE ANY OTHER TEXT REPRESENTATION APART FROM BOW?

- ▶ **N-Grams:** a continuous sequence of n items from a given sequence of text.
- ▶ Input Text: **search engines are great.**
- ▶ **BoW:** ["are", "engines", "great", "search"]
- ▶ **2-gram:** ["<bos>_search", "search_engine", "engines_are", "are_great", "great_<eos>"]
- ▶ **3-gram:** ["<bos>_search_engine", "search_engine_are", "engines_are_great", "are_great_<eos>"]

IS THERE ANY OTHER TEXT REPRESENTATION APART FROM BOW?

	Pros	Cons
BoW	Very very simple and widely used	Grammar and order are missing
n-grams	Capture local dependency and partial order	Largely increase vocabulary

POWER LAWS

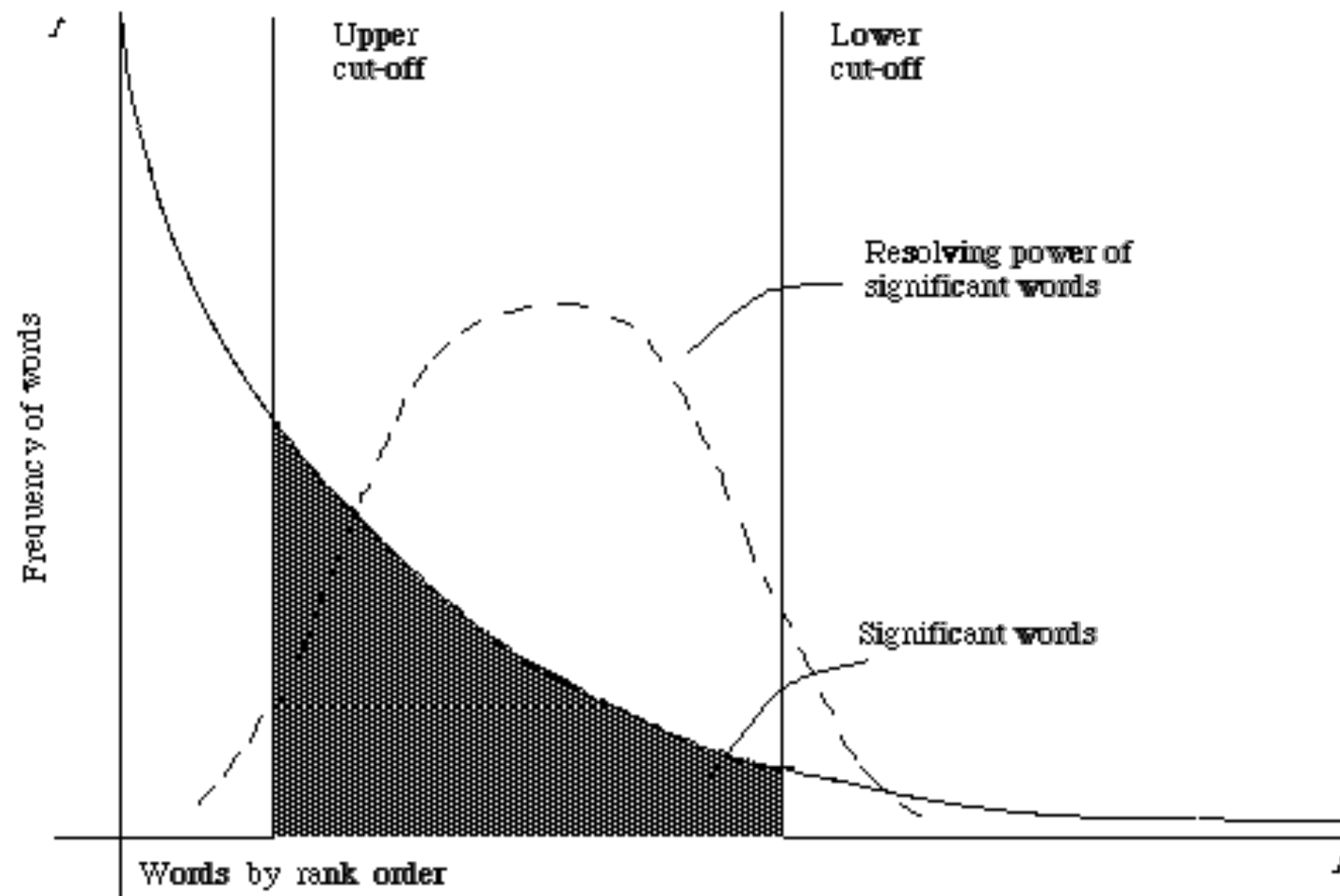


Figure 2.1. A plot of the hyperbolic curve relating f , the frequency of occurrence and r , the rank order (Adapted from Schultz⁴⁴ page 120)

LET'S MOVE TO OUR EXAMPLES WITH

PYTHON NOTEBOOK