Text Representation: Characters and Tokens

Dr. Arshin Rezazadeh

CS 4417B/9117/9647

The University of Western Ontario

Text Tasks

Text Tasks

- Information retrieval

- Summarization

- Question answering

- "Labelling" tasks

- "Structure-finding" tasks

- Annotation tasks

Information Retrieval

- Given a query, return relevant documents

- Web search
  - Google, Bing, DuckDuckGo, …

- Digital Discovery (law)

- Bespoke information retrieval: Elasticsearch

Question-answering

- Specialized information retrieval

"Hey Siri, what is the temperature outside?"
It's currently 10°C.

"Hey Siri, what is the meaning of life?"
That's easy... it's a philosophical question concerning the purpose and significance of life or existence in general.

https://huggingface.co/tasks/question-answering

Summarization

I(10): russian world no. # nikolay davydenko became the fifth withdrawal through injury or illness at the sydney international wednesday , retiring from his second round match with a foot injury .
G: tennis : davydenko pulls out of sydney with injury
A: davydenko pulls out of sydney international with foot injury
A+: russian world no. # davydenko retires at sydney international

I(11): russia 's gas and oil giant gazprom and us oil major chevron have set up a joint venture based in resource-rich northwestern siberia , the interfax news agency reported thursday quoting gazprom officials .
G: gazprom chevron set up joint venture
A: russian oil giant chevron set up siberia joint venture
A+: russia 's gazprom set up joint venture in siberia

Figure 4: Example sentence summaries produced on Gigaword. **I** is the input, **A** is ABS, and **G** is the true headline.

- https://arxiv.org/abs/१७०७.०२२६८
- https://arxiv.org/pdf/1509.00685
- https://huggingface.co/tasks/summarization

"Labelling" tasks

- Classification
  - Is an e-mail spam or not?

- Sentiment analysis
  - Guess the *emotional valence* of a piece of text. Positive? Negative? Neutral?
    - "I love my new iPhone!"
    - "I just dropped my new iPhone in the toilet. Fantastic!"

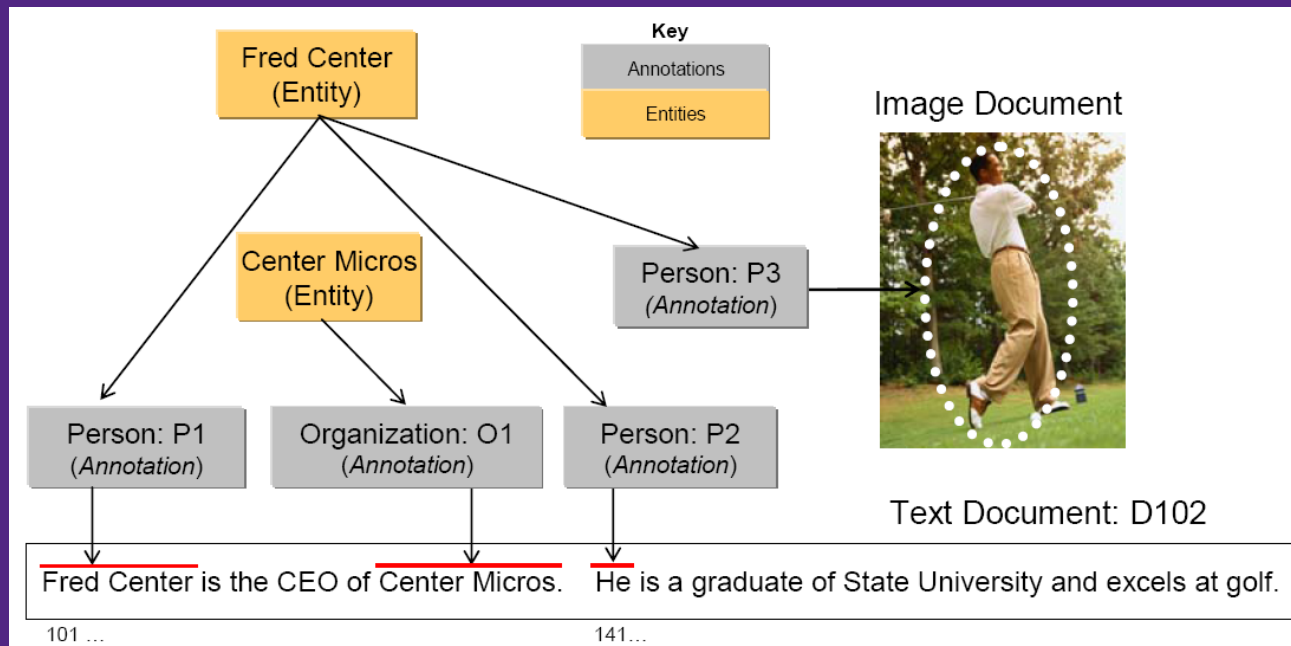- https://huggingface.co/docs/transformers/tasks/sequence_classification

"Structure-finding" tasks

- Find documents that "belong together" for some reason (clustering)
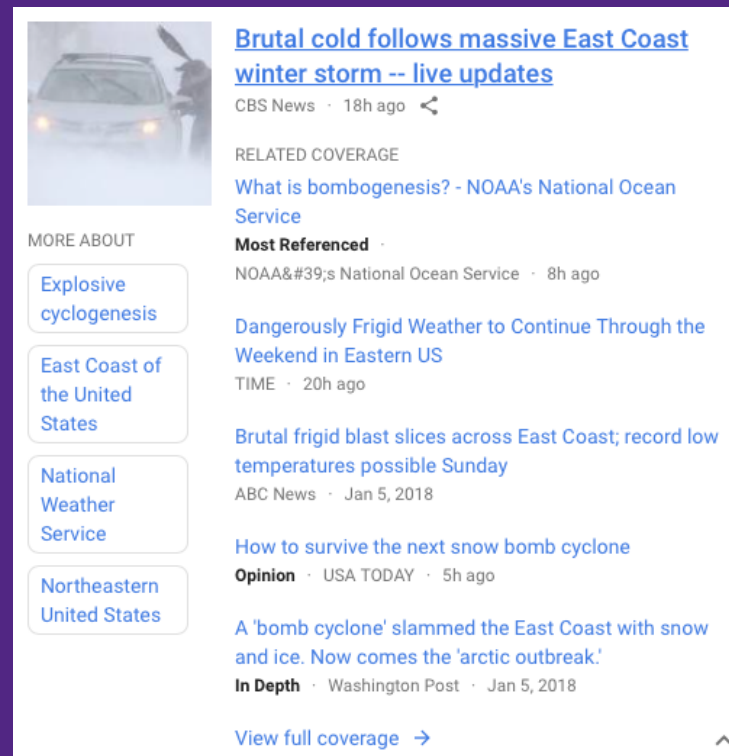
- Discover "topics" that are discussed in a collection of documents. A document may discuss more than one topic!

- https://huggingface.co/cristian-popa/bart-tl-ng

Annotation tasks

UIMA idefines building blocks called **Analysis Engines** (AEs).
One way to think about AEs is as software agents that automatically discover and record *meta-data* about original content, e.g.:
(1) The Topic of document D102 is "CEOs and Golf".
(2) The span from position 101 to 112 in document D102 denotes a Person
(3) The Person denoted by span 101 to 112 and the Person denoted by span 141 to 143 in document D102 refer to the same Entity.

Text Representation

Characters

Representations

All human beings are born free and equal in dignity and rights. They are endowed with reason and conscience and should act towards one another in a spirit of brotherhood.

41 6c 6c 20 68 75 6d 61 6e 20
62 65 69 6e 67 73 20 61 72 65
20 62 6f 72 6e 20 66 72 65 65
20 61 6e 64 20 65 71 75 61 6c
20 69 6e 20 64 69 67 6e 69 74
79 20 61 6e 64 20 72 69 67 68
74 73 2e 20 54 68 65 79 20 61
72 65 20 65 6e 64 6f 77 65 64
20 77 69 74 68 20 72 65 61 73
6f 6e 20 61 6e 64 20 63 6f 6e
73 63 69 65 6e 63 65 20 61 6e
64 20 73 68 6f 75 6c 64 20 61
63 74 20 74 6f 77 61 72 64 73
20 6f 6e 65 20 61 6e 6f 74 68
65 72 20 69 6e 20 61 20 73 70
69 72 69 74 20 6f 66 20 62 72
6f 74 68 65 72 68 6f 6f 64 2e

{'and': 4, 'are': 2, 'in': 2, 'another': 1, 'endowed': 1, 'one': 1, 'born': 1, 'beings': 1, 'human': 1, 'conscience': 1, 'should': 1, 'they': 1, 'towards': 1, 'free': 1, 'reason': 1, 'brotherhood': 1, 'with': 1, 'spirit': 1, 'a': 1, 'dignity': 1, 'all': 1, 'rights': 1, 'of': 1, 'equal': 1, 'act': 1}

Thinking Question

- What kind of problems can you already attempt with a bag of words representation?

- What kind of tasks can't be done, or require a different representation?

Character Encodings

- All documents are collections of bytes
  - 48 65 6C 6C 6F 20 57 6F 72 6C 64

- ASCII uses 1 byte per character, with values 00 through 7F. (So 7 bits.)
  - Hello World

- Extended versions like windows-1252 and ISO-8859-1 are one-byte-per-character and use codes 80 through FF for special characters

Multi-byte encodings

Hex -encoding

- UTF-8: Up to four bytes per character. Codes 00 through 7F are same as ASCII
  - 48 65 6C 6C 6F 20 57 6F 72 6C 64
  - Hello World

- Some characters use more than one byte
  - 43 68 69 6C 64 E2 80 99 73 20 50 6C 61 79
  - Child's Play

- Same bytes above, interpreted as windows-1252
  - Childâ€™s Play

Consequences

- Majority of text data on the web today are UTF-8

- "Duplicated" symbols: ′ ' ' `
  - Should these be considered the same? Different?

- Emojis!
  - 😄 😃 😊 😌 😐 🙁 😟 😱

- It's important to understand how your data are encoded, and how your software interprets the data as characters.

Naïve Search

- To search for all instances of the string "potato" 🥔 in a *corpus*, we could use a *string matching* algorithm

  征料库.

- Let n be the number of characters in the corpus
  - Even "good" algorithms are $\Omega(n)$ time
  - Google indexes at least 100 million GB worth of pages
  - Suppose CPU can examine a character in 1 nanosecond
  - Each search would take 3.17 CPU years
  - Google serves at least 5 billion searches per day...

- How do we make this faster?

Text Representation

Tokenization
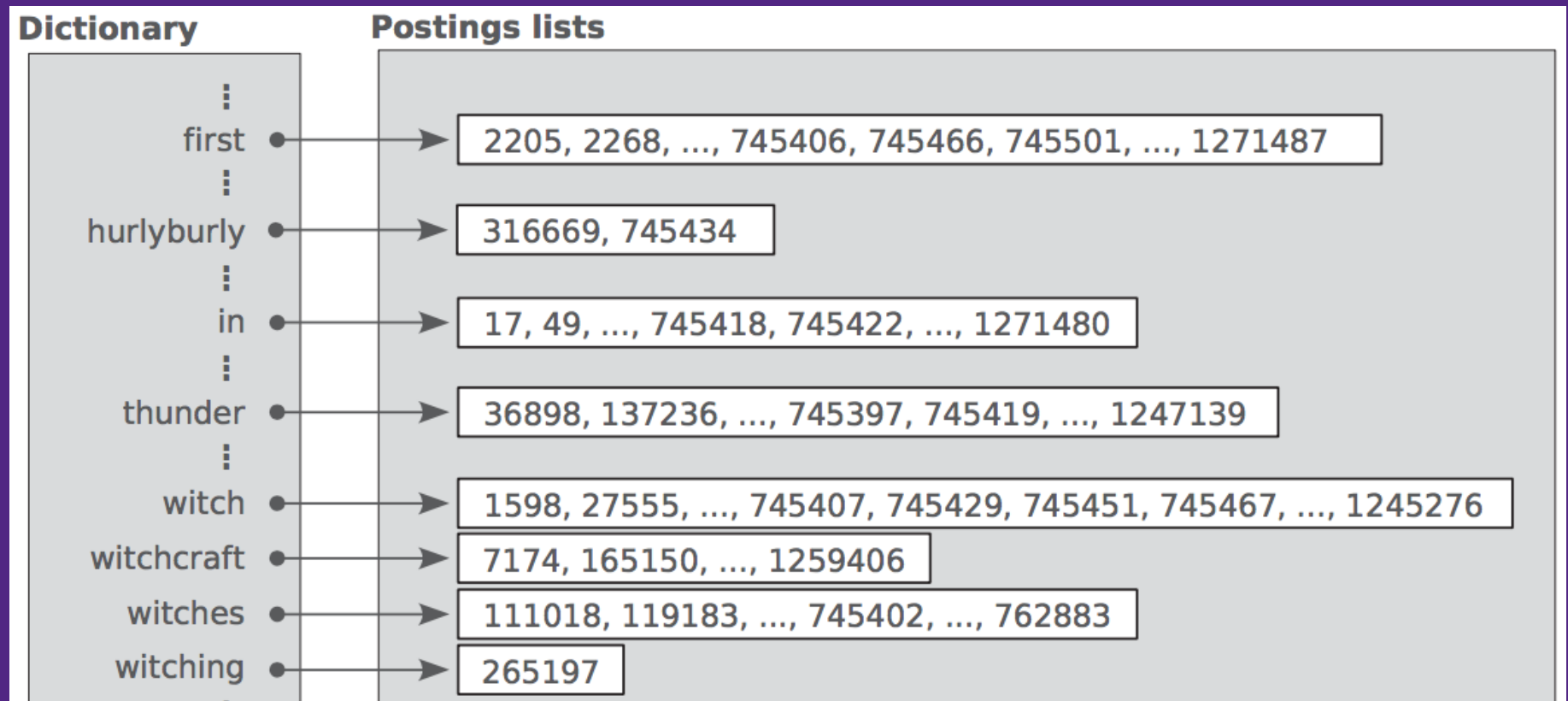
Tokens: Imposing structure

Come, gentle night, come loving, black-browed night

- A *lexical token* (or term) is a substring in a corpus
- Possible tokens from the corpus above:
    - gentle
    - black-browed
    - black
- Tokens typically selected to carry *semantic meaning*
- A collection of unique tokens is a *dictionary*
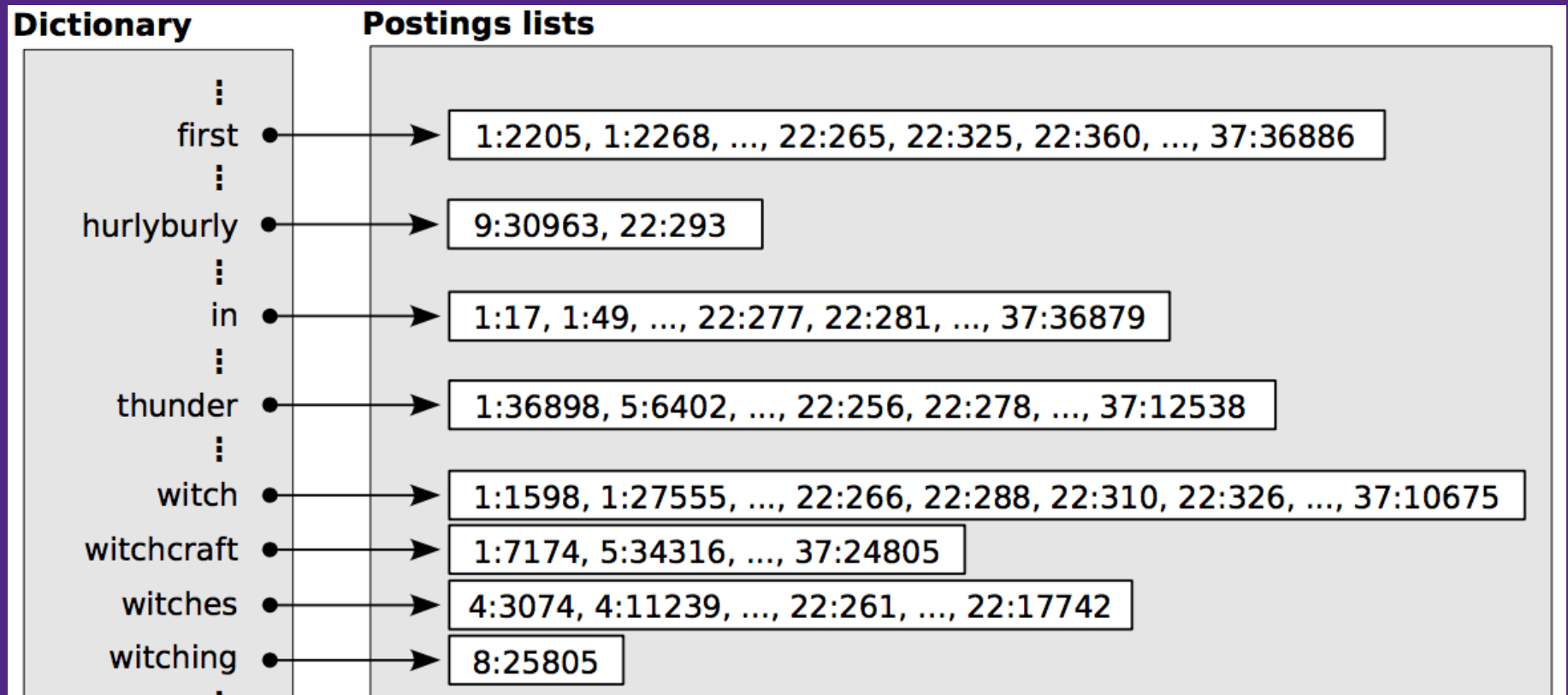
Inverted Indices

- An *inverted index* is a data structure, created from a corpus

- Given a token, tells where that token occurs in the corpus

- "Opposite" of indexing an array.
  - words[5] = "foo"
  - inverted_index(words, "foo") = 5

- Just like the index of a (paper) book

An Inverted Index

Dictionary	Postings lists
⋮	
first	2205, 2268, ..., 745406, 745466, 745501, ..., 1271487
⋮	
hurlyburly	316669, 745434
⋮	
in	17, 49, ..., 745418, 745422, ..., 1271480
⋮	
thunder	36898, 137236, ..., 745397, 745419, ..., 1247139
⋮	
witch	1598, 27555, ..., 745407, 745429, 745451, 745467, ..., 1245276
witchcraft	7174, 165150, ..., 1259406
witches	111018, 119183, ..., 745402, ..., 762883
witching	265197

How much work does it take to find the first occurrence of **night** in the corpus?

Büttcher, Clark, Cormack Ch. 2

An Inverted Index of Documents

Dictionary

Postings lists

first → 1:2205, 1:2268, ..., 22:265, 22:325, 22:360, ..., 37:36886

hurlyburly → 9:30963, 22:293

in → 1:17, 1:49, ..., 22:277, 22:281, ..., 37:36879

thunder → 1:36898, 5:6402, ..., 22:256, 22:278, ..., 37:12538

witch → 1:1598, 1:27555, ..., 22:266, 22:288, 22:310, 22:326, ..., 37:10675

witchcraft → 1:7174, 5:34316, ..., 37:24805

witches → 4:3074, 4:11239, ..., 22:261, ..., 22:17742

witching → 8:25805

Tokenization

- Roughly: identifies the *words* in documents
  - Form of annotation

- Once complete, a document is represented *only* by the sequence of tokens it generates

- Searches are done using tokens rather than original document, so **useful to map different inputs to the same token if they "mean the same thing."**

- **Tokenization is a way to explicitly ignore differences we don't care about.**

Simple Tokenization

- Early IR systems:
  - any sequence of *alphanumeric* characters of length 3 or more terminated by a space or similar (end of line, tab, …)
  - upper-case changed to lower-case

- Example:
  - "Bigcorp's 2007 bi-annual report showed profits rose 10%."
  - bigcorp 2007 annual report showed profits rose

- Small decisions in tokenizing can have major impact on information that is retained

- Information is "lost":
  - Canada, CANada, CANADA all become canada

Tokenizing Problems

- Small words can be important in some queries, usually in combinations
  - xp, pm, ben e king, el paso, master p, gm, j lo, world war II

- Sometimes hyphens should be considered part of the word, sometimes a word separator
  - winston-salem, mazda rx-7, e-cards, pre-diabetes, t-mobile, spanish-speaking, e-bay, wal-mart, active-x, cd-rom, t-shirts

Tokenizing Problems

- Special characters are an important part of tags, URLs, code in documents
  - http://

- Capitalized words can have different meaning from lower case words
  - Bush, Apple

- Apostrophes can be a part of a word, a part of a possessive, or just a mistake
  - rosie o'donnell, can't, don't, 80's, 1890's, men's straw hats, master's degree, england's ten largest cities, shriner's

Tokenizing Problems

- Numbers can be important, including decimals
  - nokia 3250, top 10 courses, united 93, quicktime 6.5 pro, 92.3 the beat, 288358

- Periods can occur in numbers, abbreviations, URLs, ends of sentences, and other situations
  - I.B.M., Ph.D., cs.umass.edu, F.E.A.R.

- **Tokenizing steps must remain consistent across documents and queries** (if applicable) for results to make sense

Tokenizing Process (Roughly)

- First, identify parts of a document that are "text" (e.g. within a web page – parse the HTML.)

- Perform **text segmentation**: Word is any sequence of alphanumeric characters, terminated by space or special character

  i.e. ([a-z][A-Z])* (space or spec char)

- Convert to lower-case

- Additional hand-written rules:
  - Ignore postrophes within words?
    - Don't -> dont
  - Ignore periods in abbreviations?
    - I.B.M. -> ibm

Unicode Text Segmentation

- Unicode Standard Annex #29

"This annex describes guidelines for determining default segmentation boundaries between certain significant text elements: grapheme clusters ("user-perceived characters"), words, and sentences. For line boundaries, see [UAX14] ."

http://unicode.org/reports/tr29/

Thinking Question

- What are some applications or domains where a "default" tokenizer would need to be modified to work well?

- Do you have any thoughts on how it should be modified?

Summary

- Text Tasks
    - Information Retrieval
    - Summarization
    - Q&A
    - Labelling
    - Structure-finding, e.g. Clustering and Topic Modelling
    - Annotation

- Representation
    - Characters
    - Tokens