NLP

SICSS-NDSU

https://slides.com/zoltanpm/sicss-ndsu-nlp/

OVERVIEW

Why is natural language so hard?

POS and dependency tagging

Basic sentiment analysis

UNSTRUCTURED TEXT, UNSTRUCTURED MEANING

SUN

SIGNIFIERS

The **material stuff** of language

SIGNIFIEDS

The **concepts** or **ideas** attached to them

EVEN MORE SIGNIFIEDS!

The weight of **cultural meanings** attached to *them*

”

Meaning can be encoded by word vector embeddings

…

but before that come simple bag-of-words models.

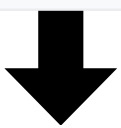
(1) John likes to watch movies. Mary likes movies too.

(2) Mary also likes to watch football games.

```
"John","likes","to","watch","movies","Mary","likes","movies","too"

"Mary","also","likes","to","watch","football","games"
```

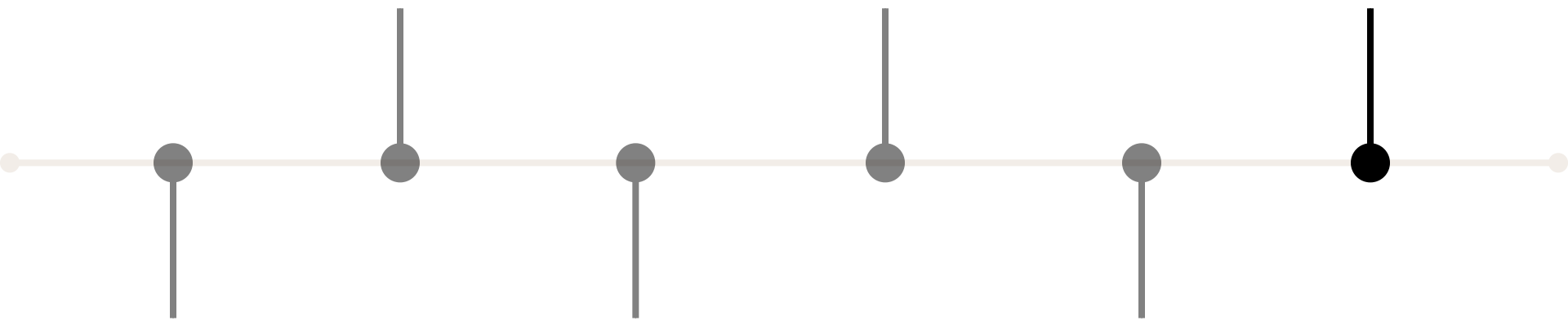
```
BoW1 = {"John":1,"likes":2,"to":1,"watch":1,"movies":2,"Mary":1,"too":1};
BoW2 = {"Mary":1,"also":1,"likes":1,"to":1,"watch":1,"football":1,"games":1};
```



Feature frequency:

```
(1) [1, 2, 1, 1, 2, 1, 1, 0, 0, 0]
(2) [0, 1, 1, 1, 0, 1, 0, 1, 1, 1]
```

BUILD A CORPUS



1

Tokenize your text

2

Make decisions about what linguistic characteristics are **relevant** to your RQs

3

Using #2, **clean text**: remove stopwords, punctuation, lowercase, etc. ... or don't.

4

Create **vocab** of words in corpus ("bag of words")

5

Convert vocab tokens to integers (or embeddings)

6

Analyze or train then analyze

TAGGING

POS	Dependency	NER
NOUN	SUBJ	PER
PROPN	OBJ	ORG
ADJ	ROOT	DATE
ADV	NMOD	GPE
https://universaldependencies.org/u/pos/	https://downloads.cs.stanford.edu/nlp/software/dependencies_manual.pdf	https://spacy.io/usage/linguistic-features

”

What's the point of things like grammatical dependency tagging for non-linguists?

What CSS questions can we answer?

OTHER CORPUS CHARACTERISTICS

- term-frequency matrix
- concordance/kwic
- keyness

  Scott (1997): ‘a word which occurs with unusual frequency in a given text [...] by comparison with a reference corpus of some kind’
- semantic similarity/(path-)distance
- dictionary-based sentiment