Semantic Text Similarity

Applications

→ Group similar words in semantic contexts
→ Textual entailment
→ Paraphrasing

Wordnet → symantic dictionary → interlinked by semantic relations.
→ organizes information in a hierarchy

Path Similarity → find shortest path between two concepts.

Lowest Common Subsumer (LCS)

→ find the closest ancestor to both concepts.

Lin Similarity → similarity measure based on the information contained in the LCS of two concepts.

$$\text{LinSim}(u,v) = \frac{2 \times \log P(\text{LCS}(u,v))}{(\log P(u) + \log P(v))}$$

$P(u)$ is given by the information content learnt over a large corpus.

Python

```
from nltk.corpus import wordnet as wn

deer.path_similarity(elk)
```

Collocations & Distributional Similarity

Two words that frequently appears in similar contexts are more likely to be semantically related.

Context → words before, words after; within a small window

→ POS words before, after, within a small window.

→ specific syntactic relation to the target word.

→ frequency of two or more words.

→ frequency of individual words.

→ so, normalisation is important.

Pointwise Mutual Information.

$$PMI(w, c) = \frac{\log P(w, c)}{P(w) \times P(c)}$$

$w \rightarrow$ word
$c \rightarrow$ context.

Python — NLTK collocations & associations
.pmi( )

Topic Modeling

Intuition

Documents are a mixture of topics.

Ex
- genetics
- computation
- life sciences
- anatomy

Topic modeling → coarse level analysis of what's in a text collection

→ topics are represented by a word distribution

→ document is assumed to be a mixture of topics.

What's known → text collection
number of topics

what's not known → actual topics
topic distribution

→ text clustering problem
→ documents + words are clustered simultaneously

Topic modeling approaches

1. PLSA (1999)
2. LDA (2003) (better)

Generative Models and LDA

Generative Models

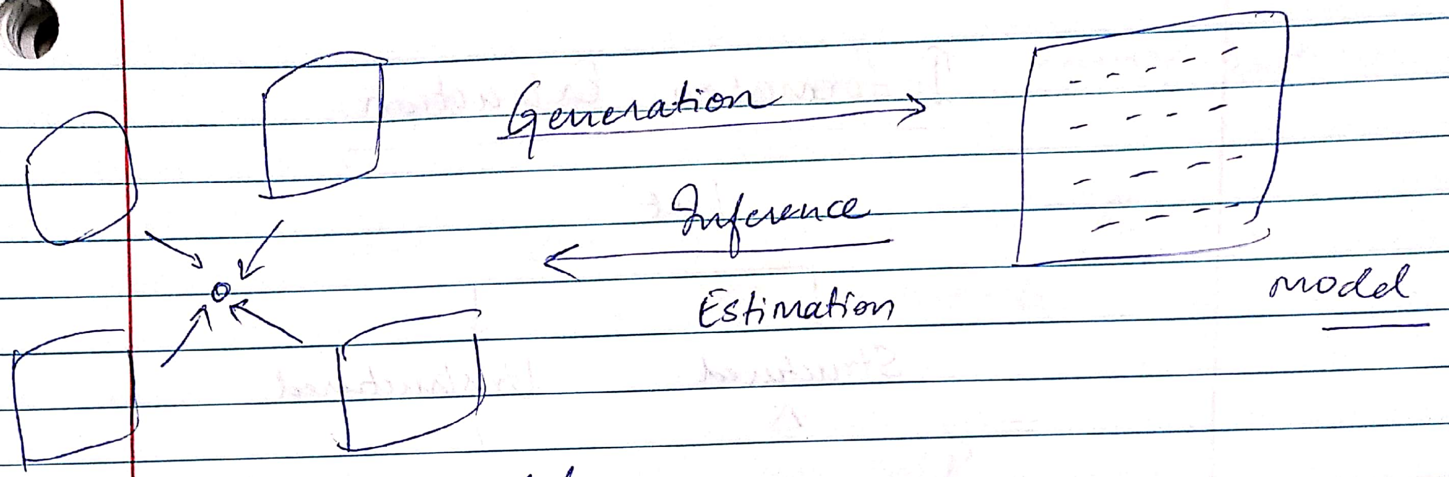
is
the harry
potter the I
movie
am

$\xrightarrow{\text{Generation}}$

the movie
harry potter
is - - -
- - - - - - - -

$\xleftarrow{\text{Inference}}$
Estimation.

Pr(text | model)



mixture model.

LDA → generative model of document d
→ mixture of topics.
→ use a topic's multinomial distribution to output words to fill that topic's quota.

In practice → How many topics?
→ Interpreting topics
→ Topics are just word distributions
→ making sense is non-trivial and subjective.

Gensim → LDA.

Preprocess → stemming, normalize, stopword removal, convert to a dtm → document term matrix.

Information Extraction

Text

Structured Unstructured

how?

Goal → identify and extracts fields of interest from free text

headline, author, reviewer, date/place of publishing, etc.

Fields of Interest

→ Named Entities. (NEWS)
  → Money, companies (finance)
  → Diseases, drugs, procedures (Medicine)

→ Relations (what, who, when, where)

NER

technique to identify all mentions.

→ identify the mention/phrase: boundary detection

→ identify the type: tagging

Approaches to identify named entities.

→ dates, phone numbers → Regex.
→ others → ML approach.

Standard NER Task in NLP

→ four class model
  → PER
  → ORG
  → LOC / GPE
  → other / outside
→ Relation extraction
→ Co-reference resolution.
  - disambiguate mentions and group mentions together.

Anita   Joseph

Her   He

This is important Q&A's.

Given a question, find the appropriate answer.