CZ4045 Natural Language Processing

Tutorial 8: Semantic Analysis

NANYANG TECHNOLOGICAL UNIVERSITY | SINGAPORE

Q1.

- Consider the example sentences in lecture notes:

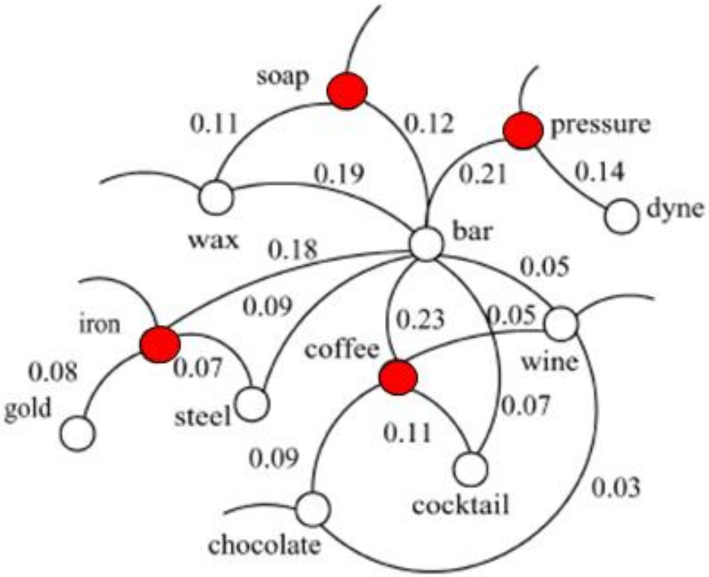
a) Use a dictionary to check the ranking position of the correct sense for the highlighted word.

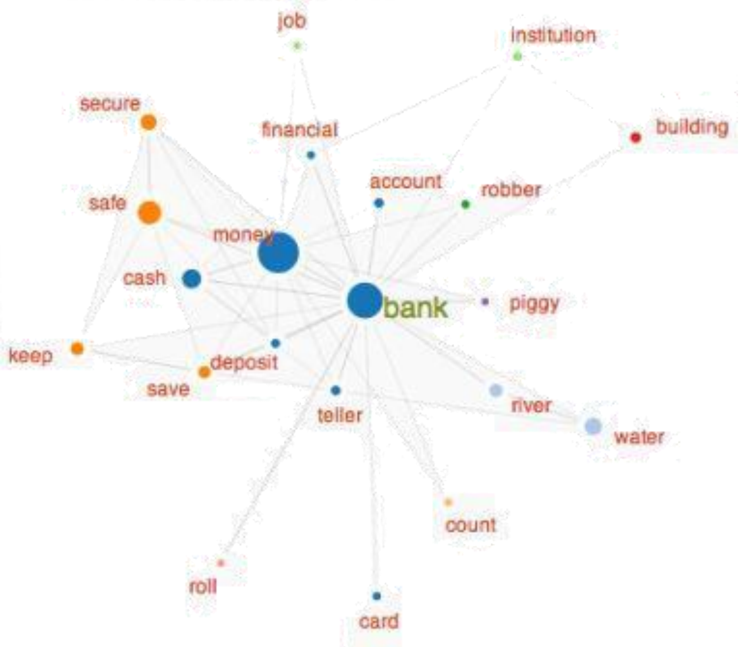
b) Check the ranking position of the correct synset in WordNet http://wordnetweb.princeton.edu/perl/webwn

c) Consider all the words in each synset form a small document. Now we can compute the similarity between the above example sentence and the small document, using cosine similarity. Is it possible to determine the right sense for the highlighted words?

Review: Word Senses

- One of the biggest challenges of NLP is the ambiguity of natural language, e.g., the same word can have many different meanings (senses) depending on the context





Review: Word Sense Disambiguation (WSD)

- The task of determining which of various senses of a word is invoked in context

- Generally viewed as a categorization task
  - Similar to POS tagging
  - Refer to a particular existing sense repository (e.g. WordNet)

- Alternative view, dividing the usages of a word into different meanings without respect to sense repository
  - Involves unsupervised techniques

Review: WSD Approaches

- Dictionary-based approach
  - use the first sense in a dictionary

- Frequency-based approach
  - choose the most frequent sense: P(sense|word)

- Supervised approach
  - train a classifier (e.g. Naïve Bayes, Support Vector Machine, or Maximum Entropy) to assign the correct sense (bag-of-words model)

Q1(a)

- Example dictionary (online) Concise English Dictionary
  - It's my **right** to do as I wish with my own body. (first)
  - The sign on the **right** was bent. (third)

  - The **plant** is producing far too little to sustain its operation for more than a year. (first)
  - An overboundance of oxygen was produced by the **plant** in the third week of the study. (second)

  - The **tank** is full of soldiers. (first)
  - The **tank** is full of nitrogen. (second)

Q1(b)

- Check the ranking position of the correct synset in WordNet
  - It's my **right** to do as I wish with my own body. (first)
  - The sign on the **right** was bent. (third)
  - http://wordnetweb.princeton.edu/perl/webwn?s=right

  - The **plant** is producing far too little to sustain its operation for more than a year. (first)
  - An overboundance of oxygen was produced by the **plant** in the third week of the study. (second)
  - http://wordnetweb.princeton.edu/perl/webwn?s=plant

  - The **tank** is full of soldiers. (first)
  - The **tank** is full of nitrogen. (second)
  - http://wordnetweb.princeton.edu/perl/webwn?s=tank

Q1(c)

- Consider all the words in each synset form a small document. Now we can compute the similarity between the above example sentence and the small document, using cosine similarity. Is it possible to determine the right sense for the highlighted words?

- Cosine Similarity

Cosine (Query, Document)

Dot product

Unit vectors

$$\cos(\vec{q},\vec{d}) = \frac{\vec{q}\bullet\vec{d}}{|\vec{q}||\vec{d}|} = \frac{\vec{q}}{|\vec{q}|}\bullet\frac{\vec{d}}{|\vec{d}|} = \frac{\sum_{i=1}^{|V|} q_i d_i}{\sqrt{\sum_{i=1}^{|V|} q_i^2}\sqrt{\sum_{i=1}^{|V|} d_i^2}}$$

q_i is the tf-idf weight of term *i* in the query

d_i is the tf-idf weight of term *i* in the document

$\cos(\vec{q},\vec{d})$ is the cosine similarity of $\vec{q}$ and $\vec{d}$ … or,
equivalently, the cosine of the angle between $\vec{q}$ and $\vec{d}$.

Length normalization

- A vector can be (length-) normalized by dividing each of its components by its length – for this we use the L2 norm:

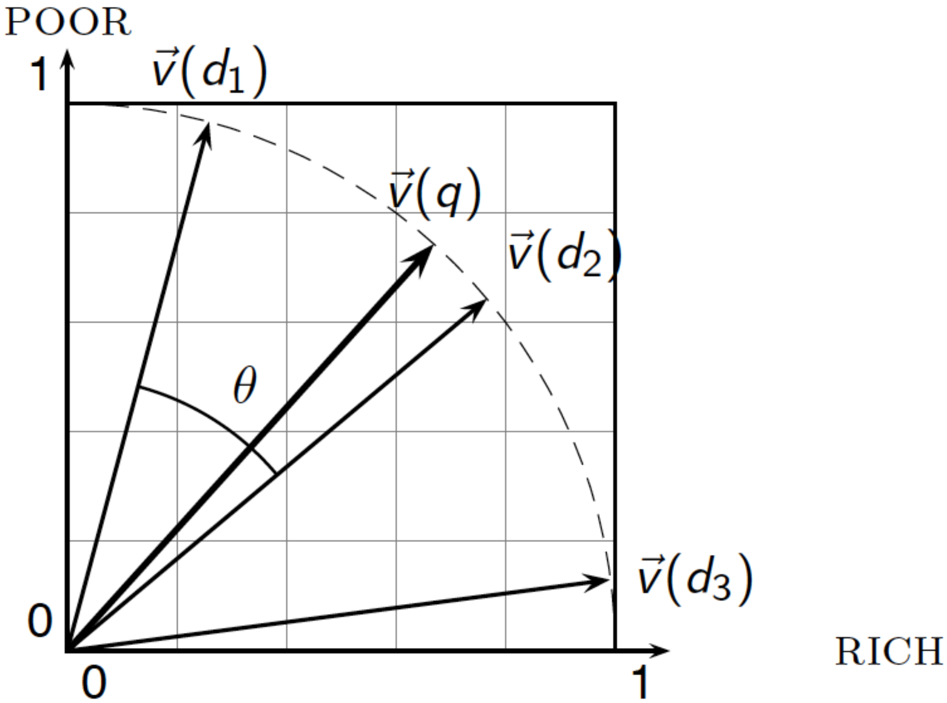
$$\left\|\vec{x}\right\|_2 = \sqrt{\sum_i x_i^2}$$

- Dividing a vector by its L2 norm makes it a unit (length) vector (on surface of unit hypersphere). Effect on the two documents d and d′ (d appended to itself) from earlier slide:
  - They have identical vectors after length-normalization.
  - Long and short documents now have comparable weights

Cosine for length-normalized vectors

- For **length-normalized** vectors, cosine similarity is simply the dot product (or scalar product):

$$\cos(\vec{q},\vec{d}) = \vec{q} \bullet \vec{d} = \sum_{i=1}^{|V|} q_i d_i$$



Cosine similarity amongst three documents: SaS, PaP, WH

term	SaS	PaP	WH
affection	115	58	20
jealous	10	7	11
gossip	2	0	6
wuthering	0	0	38

- How similar are the novels
  - SaS: *Sense and Sensibility*
  - PaP: *Pride and Prejudice*, and
  - WH: *Wuthering Heights*?

Term frequencies (counts)

Note: To simplify this example, we don't do idf weighting.

Cosine similarity amongst three documents: SaS, PaP, WH

Log frequency weighting

term	SaS	PaP	WH
affection	3.06	2.76	2.30
jealous	2.00	1.85	2.04
gossip	1.30	0	1.78
wuthering	0	0	2.58

After length normalization

term	SaS	PaP	WH
affection	0.789	0.832	0.524
jealous	0.515	0.555	0.465
gossip	0.335	0	0.405
wuthering	0	0	0.588

$\cos(\text{SaS},\text{PaP}) \approx 0.789 \times 0.832 + 0.515 \times 0.555 + 0.335 \times 0.0 + 0.0 \times 0.0 \approx 0.94$

$\cos(\text{SaS},\text{WH}) \approx 0.79$

$\cos(\text{PaP},\text{WH}) \approx 0.69$

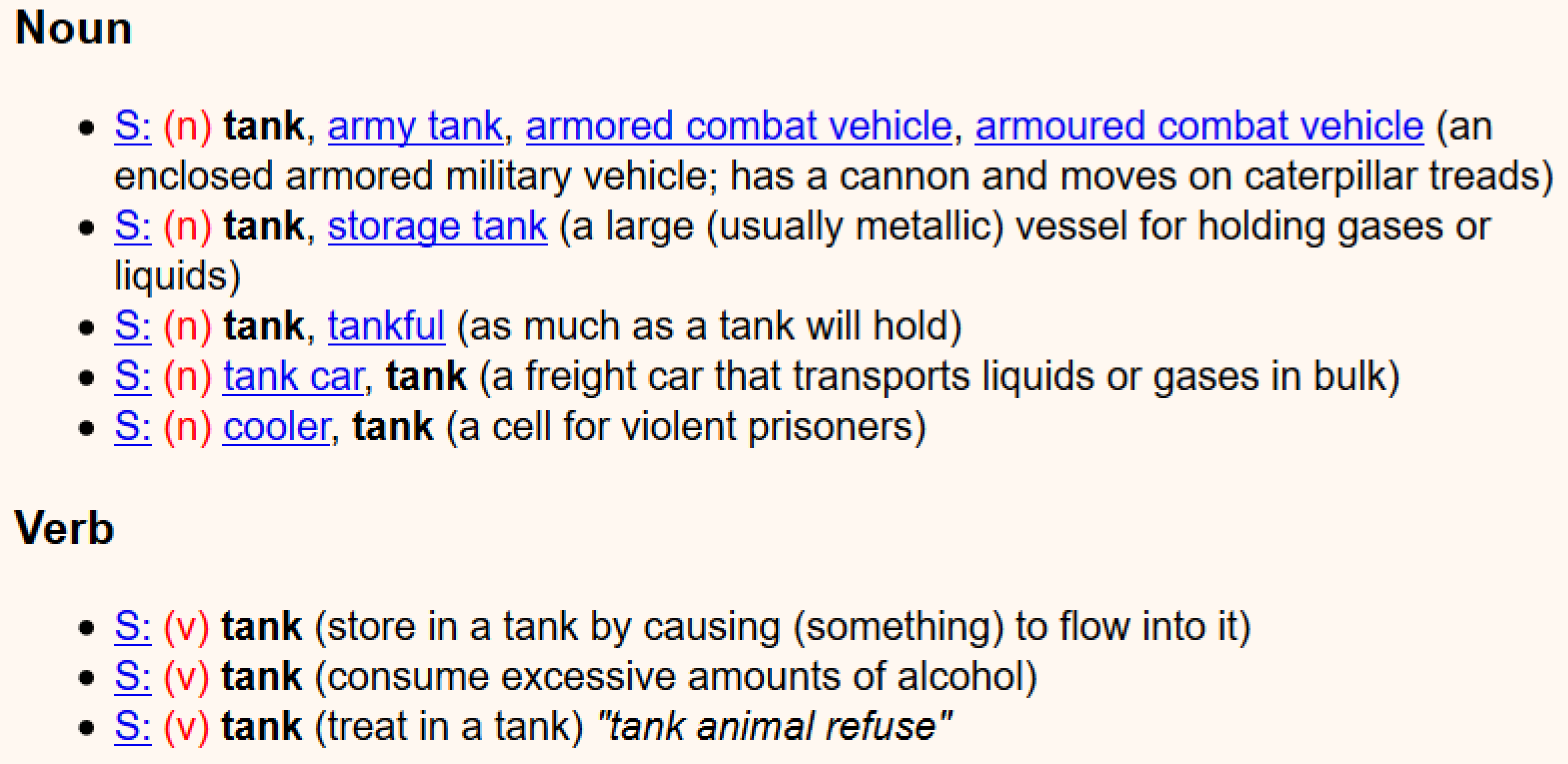
Q1(c)

- Consider all the words in each synset form a small document. Now we can compute the similarity between the above example sentence and the small document, using cosine similarity. Is it possible to determine the right sense for the highlighted words?

- Cosine Similarity
  - Pretty much depends on the degree of overlaps given both the sentence and the synset is short (in number of words).

Cosine Similarity (Synset, Sentence)

- Tank
  - The **tank** is full of soldiers. (first)
  - The **tank** is full of nitrogen. (second)



Noun

- S: (n) **tank**, army tank, armored combat vehicle, armoured combat vehicle (an enclosed armored military vehicle; has a cannon and moves on caterpillar treads)
- S: (n) **tank**, storage tank (a large (usually metallic) vessel for holding gases or liquids)
- S: (n) **tank**, tankful (as much as a tank will hold)
- S: (n) tank car, **tank** (a freight car that transports liquids or gases in bulk)
- S: (n) cooler, **tank** (a cell for violent prisoners)

Verb

- S: (v) **tank** (store in a tank by causing (something) to flow into it)
- S: (v) **tank** (consume excessive amounts of alcohol)
- S: (v) **tank** (treat in a tank) *"tank animal refuse"*

Q2: Rocchio classifier

- Each document is represented by using the bag-of-words model and the words are weighted by TF-IDF scheme.
- For each category, the centroid of category is computed as the average vector of all the documents in that category.
- To classify a new document, the cosine similarity between this document and all centroids are computed and the document is classified to the category with highest similarity.

Q2: Rocchio classifier

- Consider we have two categories and their centroid vectors are as follows:
  - ***Category 1***: human 0.23, right 0.45, law 0.3, own 0.2, wish 0.2
  - ***Category 2***: sign 0.5, position 0.6, left 0.3

- Use random values for IDFs of the words and classify the following two sentences to the two categories using Rocchio classifier:
  - *My **right** to do as I wish with my own body.*
  - *The sign on the **right** was bent.*

- Assume random values for IDF
  - D1: my 0.6 right 0.4 to 0.1 do 0.1 as 0.2 I 0.2 wish 0.5 with 0.2 own 0.3 body 0.3
  - D2: the 0.1 sign 0.4 on 0.1 right 0.4 was 0.2 bent 0.6

Q2: Rocchio classifier

- Length Normalization
  - ***C1***: human 0.35, right 0.69, law 0.46, own 0.31, wish 0.31
  - ***C2***: sign 0.60, position 0.72, left 0.36

  - D1: my 0.57 right 0.38 to 0.10 do 0.10 as 0.19 I 0.19 wish 0.47 with 0.19 own 0.29 body 0.29
  - D2: the 0.12 sign 0.46 on 0.12 right 0.46 was 0.23 bent 0.70

- Classification by cosine similarity
  - D1 → C1:0.35 (right, own) C2: 0
  - D2 → C1:0.32 (right) C2: 0.276 (sign)