

From frequency to meaning: vector space models of semantics

CIS 530 Lecture 2

Measuring similarity between objects

- Many tasks can be expressed as a question of similarity
 - Information retrieval: the search engine should return documents that are similar to the user query
 - Document classification
 - Clustering of objects: find groups of similar items
 - Document segmentation
 - Essay grading
 - SAT/GRE language questions

Lecture goal for today

- Introduce a simple yet powerful way of representing text, words etc
- Introduce several similarity metrics that work well with such representations

Spatial analogy

- Think about points (locations) in this room
- Choose a coordinate system
 - Each point is represented by a vector (d_1, d_2, d_3)
 - We can compute distance between points (i.e. Euclidean distance)

$$d(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + (p_3 - q_3)^2}$$

- Points that are near each other are similar

Think about representing document

- Choose a list of words $w_1, w_2, w_3, \dots, w_n$
 - These could be for example all the words that appear in a collection of documents
- Each of these words corresponds to one axis of a coordinate system
 - The corresponding vector representation entry can be for example the number of times a word occurs

Representing three documents

- S1: The key is in the backpack.
S2: The key is in the front pocket of the backpack.
S3: The bear den is in the far end of the forest.

| | the | key | is | in | backpack | front | pocket | of | bear | de n | far | end | forest |
|----|-----|-----|----|----|----------|-------|--------|----|------|---------|-----|-----|--------|
| S1 | 2 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| S2 | 3 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| S3 | 3 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 |

$$\text{dist}(S_1, S_2) = \sqrt{4}$$

$$\text{dist}(S_1, S_3) = \sqrt{9}$$

$$\text{dist}(S_2, S_3) = \sqrt{9}$$

Completing the vector representation

- There are a number of reasonable choices about what the weight of each word should be
- Binary: 1 if present 0 if not
- Number of occurrences
- Tf.idf
- Mutual information

Tf.idf

- Tf is term frequency
 - Number of occurrences of the word in the document
- Idf is inverse document frequency $\log \frac{N}{df}$
 - Consider a collection of N documents
 - df for a particular word is the number of documents (among these N) that contain the word

idf weight

- df_t is the document frequency of t : the number of documents that contain t
 - df_t is an inverse measure of the informativeness of t
 - $df_t \leq N$
- We define the idf (inverse document frequency) of t by

$$idf_t = \log_{10} (N/df_t)$$

- We use $\log (N/df_t)$ instead of N/df_t to “dampen” the effect of idf.

Turns out the base of the log is immaterial.

idf example, suppose $N = 1$ million

| term | df_t | idf_t |
|-----------|-----------|---------|
| calpurnia | 1 | |
| animal | 100 | |
| sunday | 1,000 | |
| fly | 10,000 | |
| under | 100,000 | |
| the | 1,000,000 | |

$$idf_t = \log_{10} (N/df_t)$$

There is one idf value for each term t in a collection.

Mutual information

- Consider for example that we have several topics (sports, politics, entertainment, academic)
- We want to find words that are descriptive of a given topic
 - Use only these in our representations
 - Else small irrelevant differences can accumulate to mask real differences

$$\log \frac{p(\text{topic}_i, \text{word}_j)}{p(\text{topic}_i)p(\text{word}_j)}$$

The probabilities can be easily computed from counts in a sample collection

How?

Pointwise mutual information will have high value when there is strong relationship between the topic and the word

Measures of distance and similarity

- Euclidean distance is actually never used in language related applications
- This distance is large for vectors of different length

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

cosine similarity of q and d or, equivalently, the cosine of the angle between q and d .

Distance metrics

$$\mathit{dist}_M(p, q) = \sum_{i=1}^n |p_i - q_i|$$

$$\mathit{dist}_C(p, q) = \max_i (|p_i - q_i|)$$

$$\mathit{dist}(p, q) = \left(\sum_{i=1}^n |p_i - q_i|^N \right)^{\frac{1}{N}}$$

Similarity metrics

$$sim_J(p, q) = \frac{\sum_{i=1}^N \min(p_i, q_i)}{\sum_{i=1}^N \max(p_i, q_i)}$$

$$sim_D(p, q) = 2 \sum_{i=1}^N \min(p_i, q_i) / \sum_{i=1}^N (p_i + q_i)$$

How would to choose weighting and similarity measure

- There is no right answer
- But looking at the definitions of the metrics you could notice some possible problems and peculiarities
- Try out and see what works best for a given application
 - Get some sample results you would like to obtain
 - Try out different possibilities
 - See which one is closest in predicting the results you would like

Departures from the special analogy

- Our “dimensions” are not independent of each other
 - Most likely more than necessary
- In new examples we are likely to see new dimensions
 - Words that we did not see when we decided on vocabulary
- Highly dimensional space
 - There is a huge number of words in a language
 - So the weighting schemes are trying to deal with such issues