Assignment11 – Natural Language

Given: July 17   Due: July 22

Problem 11.1 (Ambiguity)

1. Explain the concept of *ambiguity* of *natural languages*.

Solution: *Ambiguity* is the *phenomenon* that in *natural languages* a single *utterance* can have multiple *meanings*.

---

2. Give two examples of different kinds of *ambiguity* and explain the *readings*.

Solution: Here are some examples

- "bank" can be a financial institution or a geographical feature.
- In "I saw her duck" the word "duck" can be a *verb* or a *noun*.
- "Time flies like an arrow" could be about the preferences of special insects ("time flies") or about the fact that time passes quickly – e.g. in an exam.
- In "Peter saw the man with binoculars", it could be Peter who is using binoculars, or it could be that Peter saw "the man" who had "binoculars".

Problem 11.2 (Language Identification)

You are given an English, a German, a Spanish, and a French text corpus of considerable size, and you want to build a language identification algorithm A for the EU administration. Concretely A takes a string as input and classifies it into one of the four languages $\ell^* \in \{English, German, Spanish, French\}$. The prior probability distribution for the strings being English/German/Spanish/French, is $\langle 0.4, 0.2, 0.15, 0.15\rangle$.

How would you proceed to build algorithm A? Specify the general steps and give/derive the formula for computing ℓ given a string $\mathbf{c}_{1:N}$.

Solution:

1. Build a trigram *trigram language model* $\mathbf{P}(c_i \mid \mathbf{c}_{i-2:i-1}, \ell)$ for each candidate language ℓ by counting trigrams in an ℓ*-corpus*.

2. Apply *Bayes' rule* and the *Markov property* to get the most likely language:

$$\begin{aligned}
\ell^* &= \underset{\ell}{\operatorname{argmax}}\,(P(\ell \mid \mathbf{c}_{1:N})) \\
&= \underset{\ell}{\operatorname{argmax}}\,(P(\ell) \cdot P(\mathbf{c}_{1:N} \mid \ell)) \\
&= \underset{\ell}{\operatorname{argmax}}\,(P(\ell) \cdot (\prod_{i=1}^{N} P(\mathrm{c}_i \mid \mathbf{c}_{i-2:i-1}, \ell)))
\end{aligned}$$

Problem 11.3 (Language Models)

1. How can we obtain a *trigram* model for a *natural language*?
   Explain the *probability distribution* involved.

   Solution: We need a *corpus* of words over L. Then we count how often each *trigram* occurs in it and use that to estimate the probability distribution $P(T = t)$ of trigrams t.

   ---

2. Explain informally how we can use *trigram* models to identify the language of a document D.

   Solution: We build a *trigram* model for each candidate language. Then we use each model to compute the probability of D occurring in that language. We choose the language with the highest probability.

   ---

3. Explain briefly what *named entity recognition* is.

   Solution: The task of finding, in a text, names of things and deciding what class they belong to.

Problem 11.4 (Information Retrieval)

Let D be the set containing the following three texts:

- d_1: Decision theory investigates decision problems: how an agent deals with choosing among actions.
- d_2: Reinforcement learning is a type of unsupervised learning where an agent learns how to behave in an environment.
- d_3: *Information retrieval* deals with representing information objects.

Let q be the query "agent action".

1. Give the list of words occurring in any of these texts and the word frequency $tf(t,d)$, i.e., the number of occurrences of t in d divided by the length of d (measured in words), for each text d. Normalize all words so that inflection (plural, -ing, etc.) is ignored.

Solution: The order of the list does not matter as long as it is fixed. We use

decision theory investigate problem how
a agent deal with choose
among action reinforcement learn be
type of unsupervised where to
behave in environment *information retrieval*
represent object

The word frequency vectors for the three texts are

$$tf(\_,d_1) : \langle 2,1,1,1,1, \quad 1,1,1,1,1, \quad 1,1,0,\dots,0\rangle/13$$

$$tf(\_,d_2) : \langle 0,0,0,0,1, \quad 3,1,0,0,0, \quad 0,0,1,3,1, \quad 1,1,1,1,1, \quad 1,1,1,0\dots,0\rangle/18$$

$$tf(\_,d_3) : \langle 0,0,0,0,0, \quad 0,0,1,1,0, \quad 0,0,0,0,0, \quad 0,0,0,0,0, \quad 0,0,0,2,1, \quad 1,1\rangle/7$$

---

2. For every word t, give the inverse document frequency $idf(t,D)$.

Solution: We have $N = 3$. Let $N(t) := \{d \in D \mid t \in d\}$. Then

$$N(\_) = \langle 1,1,1,1,1, \quad 2,2,2,2,1, \quad 1,1,1,1,1, \quad 1,1,1,1,1 \quad 1,1,1,1,1, \quad 1,1\rangle$$

$idf(t,D)$ is $\log_{10} 3/N(t)$: it is 0.48 for $N(t) = 1$ and 0.18 for $N(t) = 2$.

---

3. For every word t and every document, give $tfidf(t,d,D)$. Do the same for the query q "agent action".

Solution: $tfidf(t,d,D)$ is obtained by multiplying $tf(t,d)$ with 0.48 (if $N(t) = 1$) or 0.18 (if $N(t) = 2$) or 0 (if $N(t) = 3$), e.g.,

$$tfidf(\_,d_1,D) : \langle 0.96,0.48,0.48,0.48,0.48, \quad 0.18,0.18,0.18,0.18,0.48, \quad 0.48,0.48,0,\dots,0\rangle/13$$

For the query, all values are 0 except for $tfidf(\text{agent},q,D) = 1/2 \cdot 0.18$ and $tfidf(\text{action},q,D) = 1/2 \cdot 0.48$.

---

4. Compute the cosine similarity for q and each d_i.

Solution: Let $A_i = tfidf(\_,d_i,D)$ and $B = tfidf(\_,q,D)$. We have $A_1 \cdot B =$

$(1 \cdot 1 \cdot 0.18^2 + 1 \cdot 1 \cdot 0.48^2)/(13 \cdot 2)$ and $|A_1| = \sqrt{0.96^2 + 4 \cdot 0.18^2 + 7 \cdot 0.48^2}/13$ and $|B| = \sqrt{0.18^2 + 0.48^2}/2 = 0.25$. Then, we obtain $\cos\theta_1 = A_1 \cdot B/(|A_1| \cdot |B|)$. $\cos\theta_2$ and $\cos\theta_3$ are obtained accordingly.

---

5. How is the cosine similarity used to answer the query?

Solution: We return the document with the highest cosine similarity or return the list of documents ordered decreasingly by cosine similarity.
