Assignment11 – Natural Language

Given: July 6 Due: July 18

Problem 11.1 (Ambiguity) 30 pt

1. Explain the concept of ambiguity of natural languages.
2. Give two examples of different kinds of ambiguity and explain the readings.

Problem 11.2 (Language Models) 30 pt

1. How can we obtain a trigram model for a language? Explain the probability distribution involved.
2. Explain informally how we can use trigram models to identify the language of a document D.
3. Explain briefly what named entity recognition is.

Problem 11.3 (Information Retrieval) 40 pt

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.
2. For every word t, give the inverse document frequency $idf(t,D)$.
3. For every word t and every document, give $tfidf(t,d,D)$. Do the same for the query q “agent action”.
4. Compute the cosine similarity for q and each d_i.
5. How is the cosine similarity used to answer the query?