**Section 1: Semantics**

**Question 1A (10 Marks)**

Consider the following frequency vectors:

**black 4 4 2 6 0**

**white 4 8 0 2 10**

Using cosine similarity, compute the distributional semantic distance between ‘black’ and ‘white’.

**Answer:**

Black=(4,4,2,6,0)

White=(4,8,0,2,10)

Similarity=(4\*4+4\*8+0+6\*2+0)/sqrt(72\*184)=60/115.1=0.52

**Question 1B (10 Marks)**

Consider the following sense definitions for ‘bank’:

**bank#1 - the slope beside a body of water**

**bank#2 - a financial institution that accepts deposits and channels the money into lending activities**

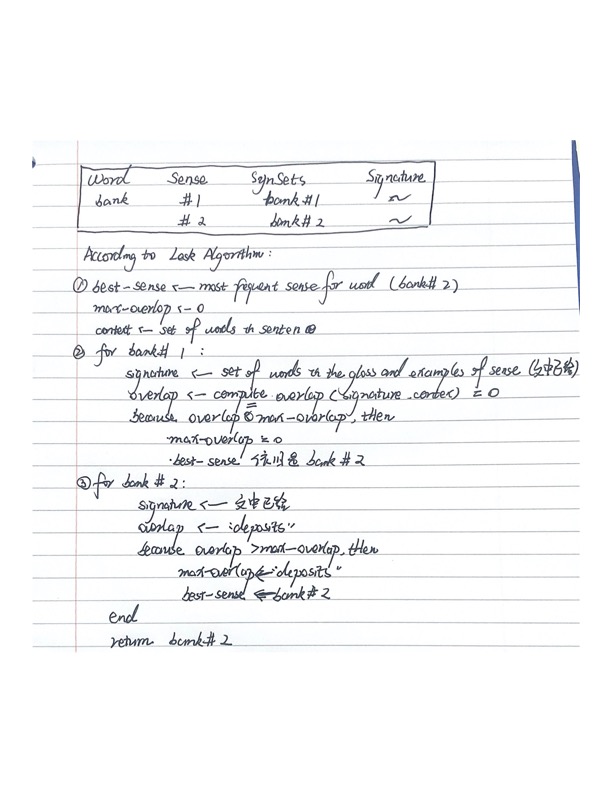
Now consider the following occurrence of ‘bank’ in this sentence:

The bank was left free to offer interest on demand deposits.

How would you apply the Lesk algorithm to disambiguate ‘bank’ between the two senses given above?

**Text

Description automatically generated**

****

**Question 1C (5 Marks)**

Define homonymy, synonymy, and antonymy.

Antonyms (反义词) are words with opposite meanings. E.g. happy-sad

Synonyms （近义）are words with the same or similar meaning. E.g. quick-fast

Homonyms (一词多义)are words that are spelled and pronounced the same, but have different meanings. E.g. chair: 1) an item of furniture; 2) the head of a department

**Section 2: Part-of-speech tagging**

**Question 2A (10 Marks)**

Consider a Hidden Markov Model with the following probabilities (S designates the

start state): What is the probability of the sequence “the University of Ireland” being tagged as “O B I I”?

**Answer:**

P= P(O|Start)\*P(B|O)\* P(I|B)\*P(I|I)\*P(the|O)\* P(university|B)\*P(of|I)\*P(Ireland|I)

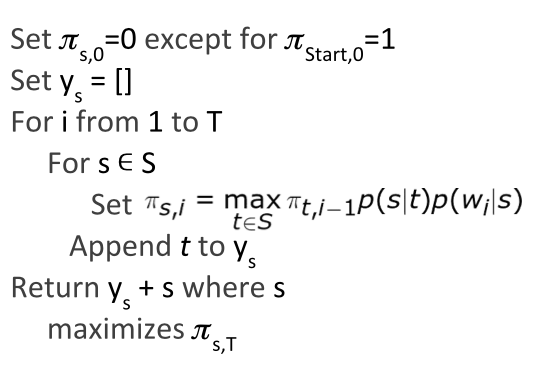
=0.4\*0.3\*0.6\*0.4\*0.6\*0.7\*0.5\*0.4=0.0024192

**Question 2B (15 Marks)**

Given an annotated text corpus, describe how you would find probabilities such as

given in the table above. Write any algorithms you would use in pseudo-code.

**Answer:** We can use **Baum-Welch algorithm** to find an optimal parameters.

****

**Besides, if we know the tags for words (supervised learning) we can directly obtain the probabilities by counting.** Diagram

Description automatically generated with medium confidenceA picture containing shape

Description automatically generated

**Section 3: Sentiment Analysis**

**Question 3A (10 Marks)**

Explain two challenges for automatic approaches to sentiment analysis.

**Answer:**

1. Implicit sentiment

Neutral words used but POS sentiment implied. E.g. “go read the book!”

Neutral words used but NEG sentiment implied. E.g. “if you are reading this because it is your darling fragrance, please wear it at home exclusively and tape the windows shut.”

1. Ambiguity

Same words used in different contexts express NEG vs. POS sentiment.

e.g. “This car’s steering is unpredictable!”

e.g. “This film is unpredictable!”

3) Irony & Sarcasm (反讽与讽刺)

Positive words used but NEG sentiment implied.

e.g. “Great job Rogers! Raise rates but not service.”

4) negation

Positive words used but NEG sentiment implied.

“I don’t like this new Nokia model.”

“I didn’t enjoy it.”

5) Informal language

Social media content uses non-standard, informal language such as hashtags.

6) indirect sentiment

Sentiment expressed may not. Be that of the author.

“Although this product is disliked by many,..”

**Question 3B (5 Marks)**

**What is a sentiment lexicon and how may it be used as a feature in a sentiment analysis classifier?**

Text

Description automatically generated

Classification task is to identify if a given text has positive or negative sentiment. Classifiers can develop on the basis of a sentiment lexicon (unsupervised) and sentiment labeled data (supervised). **Sentiment lexicon can be used to define the polarity of a word as the probability of co-occurrence with positive/negative words**.

**Question 3C (5 Marks)**

Provide a suggestion of one way in which negation may be handled in a sentiment

analysis.

**Graphical user interface, text, application

Description automatically generated**

**Question 3D (5 Marks)**

What is meant with aspect-based sentiment analysis? Give an example of an aspect.

**A picture containing text, screenshot, indoor

Description automatically generated**

**Table

Description automatically generated**

**Section 4: Information Extraction**

**Question 4A (15 Marks)**

Consider the following corpus of sentences about company acquisitions, with named entity annotation (COM:company) and gold standard labeling if the sentence does or does not express a company acquisition:

Evaluation metrics derive from Information Retrieval: Precision, Recall, F-Score Evaluation against a “Gold Standard”: human annotated (labelled) data items.

**Graphical user interface, text, application

Description automatically generated**

Table

Description automatically generatedTable

Description automatically generated

**Answer:**

#Gold standard items (GS)- 4 (1 NEG, 4 POS)

#Extracted items (EX)- 3

#Correctly extracted items (CEX) – 2

**P=CEX/EX**=2/3

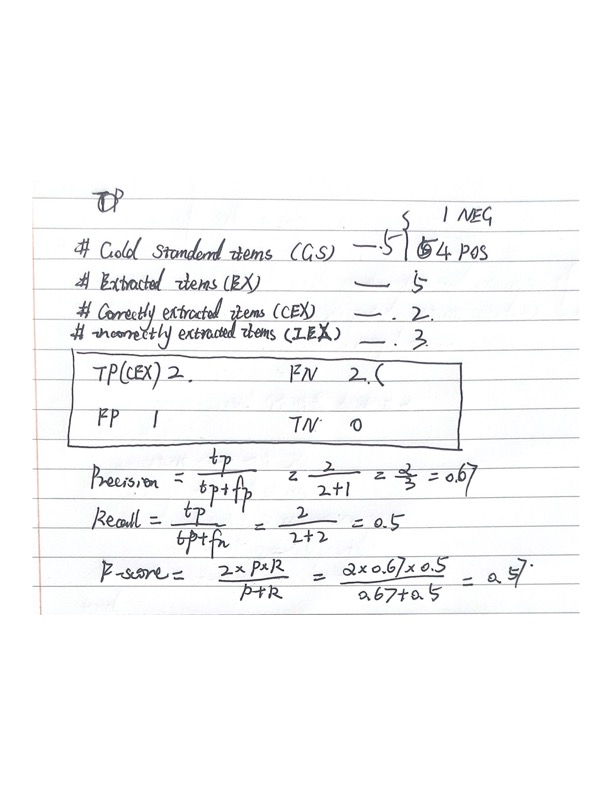
**R=CEX/GS**=2/4

**F=2\*P\*R/(P+R)**

Method 2:

**TP=CEX**=2, **FP=EX-CEX**=3-2=1, **FN=GS-CEX**=4-2=2.

Because **P=TP/(TP+FP), R=TP/(TP+FN),** Thus:



**Question 4B (10 Marks)**

Give the formula for Cohen's kappa coefficient. What is it used for in information extraction?

Graphical user interface, text, application, email

Description automatically generatedGraphical user interface, application

Description automatically generated

Table

Description automatically generated