University of Zagreb
Faculty of Electrical Engineering and Computing

April 30, 2014

Text Analysis and Retreival – Midterm Exam (AY 2013/2014)

*The exam has **20 questions** for a total of **35 points**. All multi-choice questions carry 1 point each (1/2 point subtracted for incorrect answer), while problem questions carry 4 points each. The exam duration is **120 minutes**. You must turn in the exam questions with your solutions.*

Part I: Multi-choice questions (*15 points*)

1. (*1 pt*) The Porter stemmer uses conditions imposed on suffix stripping rules. What is the general purpose of these conditions?
   (a) Prevent understemming (c) Increase coverage (e) Prevent overstemming
   (b) Remove plural suffixes (d) Reduce the number of rules

2. (*1 pt*) Select the incorrect claim about cluster information retrieval model:
   (a) Hierarchical agglomerative clustering with complete linkage produces best results on small collections
   (b) Clustered retrieval reduces efficiency at retrieval time due to additional clustering step
   (c) All documents in the cluster of the query-relevant centroid are returned
   (d) Clustered retrieval models achieve better recall of relevant documents
   (e) Hierarchical agglomerative clustering with average linkage produces best results on large collections

3. (*1 pt*) The main disadvantage of Laplace smoothing is that it
   (a) Assumes all documents are of equal length
   (b) Ignores frequencies of observed words
   (c) Assumes word order does not matter
   (d) Favors words from shorter documents
   (e) Assumes that all unseen words are equally likely

4. (*1 pt*) Which of the following is not a Topic Detection and Tracking (TDT) task?
   (a) Link detection (c) Topic detection (e) Topic summarization
   (b) Story segmentation (d) Topic tracking

5. (*1 pt*) Which of the following statements about random indexing is *not* correct?
   (a) Random indexing is a topic modeling algorithm
   (b) Random indexing is computationally less complex than latent semantic analysis
   (c) Context vectors for words are produced by summing the index vectors of all contexts in which the word appears
   (d) The dimensionality reduction in random indexing is implicit
   (e) Index vectors are high-dimensional sparse vectors with a small number of randomly distributed $+1$ and -1 values

6. (*1 pt*) Consider a collection of 10 documents and 2 user queries which must be processed on the collection. Assume the search engine uses language modeling for retrieval, as described by the query likelihood model. How many language models in total must be built to accomplish this retrieval task (assume no smoothing is used):
   (a) 10 (b) 12 (c) 13 (d) 11 (e) 1

7. (*1 pt*) You're encoding word features for your ML model using one-hot encoding. Your corpus has 10M words and 100K unique words. You've decided to discard from your corpus all words that occur twice or less. How big is your one-hot feature vector?

   (a) ∼10M dimensions (c) ∼25K dimensions (e) ∼1K dimensions

   (b) ∼100M dimensions (d) ∼100k dimensions

8. (*1 pt*) Language ambiguity is a problem for text analysis. Which of the following sentences exemplifies *categorial ambiguity*?

   (a) *"John loves Marry."*

   (b) *"Flying planes can be dangerous."*

   (c) *"Lisa gave Ann a present and she said thanks."*

   (d) *"The thief was charged by the police and had to pay a fine."*

   (e) *"I saw a boy on the hill with a telescope."*

9. (*1 pt*) The *main* advantage of the two-Poisson model over the binary relevance retrieval model is that it:

   (a) Accounts for word order (d) Accounts for word frequencies

   (b) Accounts for document length (e) Gives a relevance score which is not binary

   (c) Is more computationally efficient

10. (*1 pt*) Typical NLP tools are POS tagging (PT), lemmatization (L), sentence segmentation (SS), tokenization (T), and parsing (P). How does the typical NLP pipeline look like?

    (a) SS → T → PT → L → P (c) T → SS → P → L → PT (e) T → SS → PT → L → P

    (b) SS → T → PT → P → L (d) SS → T → P → L → PT

11. (*1 pt*) Select the valid claim regarding the HITS algorithm:

    (a) A page is a good authority if it points to good hubs

    (b) A page is a good hub if it is pointed to by other good hubs

    (c) A page is a good authority if it points to other good authorities

    (d) A page is a good hub if it points to other good hubs

    (e) A page is a good hub if it points to good authorities

12. (*1 pt*) Which of the following is a true shortcoming of the supervised machine learning models for information extraction?

    (a) Machine learning models are difficult to adapt to new domains

    (b) It is difficult to model subjectivity with machine learning models

    (c) Data labeling can be expensive and tedious

    (d) Relies on manually labeling data instead of figuring out an algorithm to solve the problem

    (e) Labeling cannot be done without expert knowledge

13. (*1 pt*) In NLP, using sequence labeling models such as HMM and CRF rather than performing sequence labeling as classification is preferred because:

    (a) HMM and CRF allow to integrate labels from both side surrounding tokens as features

    (b) There is an assumption of independence of individual classification decisions

    (c) The uncertainty of token-wise decisions is not propagated

    (d) Labels of tokens are not dependent on the labels of other tokens in the sequence

    (e) Any classification algorithm (e.g., naïve Bayes, SVM) can be pluged into a sequence labeling models such as HMM and CRF

14. (*1 pt*) You are given a collection of three documents: d_1 = "Frodo and Gandalf fought Sauron"; d_2 = "Gandalf the Gray became Gandalf the White"; d_3 = "Gray clouds circled above Frodo and Gandalf"; and the query q = (Frodo $\wedge$ Gandalf) $\vee$ (Gandalf $\wedge$ Gray). Using the inverted file index, which of the expressions retrieves the relevant documents for the query?

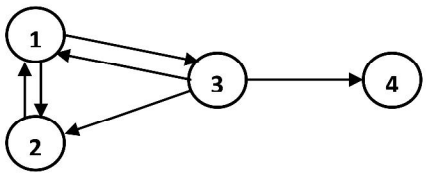
    (a) $(\{d_1, d_2, d_3\} \cap \{d_2, d_3\}) \cup (\{d_1, d_3\} \cap \{d_1, d_2, d_3\})$
    (b) $(\{d_2, d_3\} \cap \{d_1, d_3\}) \cup (\{d_2, d_3\} \cap \{d_1, d_2\})$
    (c) $(\{d_1, d_2, d_3\} \cap \{d_3\}) \cup (\{d_1\} \cap \{d_1, d_2, d_3\})$
    (d) $(\{d_1, d_3\} \cup \{d_1, d_2, d_3\}) \cup (\{d_1, d_2, d_3\} \cup \{d_2, d_3\})$
    (e) $(\{d_1, d_2, d_3\} \cap \{d_2, d_3\}) \cap (\{d_1, d_3\} \cap \{d_1, d_2, d_3\})$

15. (*1 pt*) In the Latent Dirichlet Allocation generative story, we draw $w_{d,n}$ (the n-th word of the d-th document) from a particular topic ϕ_i that is:

    (a) Most likely in document d
    (b) Most likely given $w_{d,n-1}$
    (c) Defined by the corresponding $z_{d,n}$
    (d) Drawn from θ_d
    (e) A Dirichlet distribution

Part II: Problem questions (*20 points*)

16. (*4 pts*) The PageRank algorithm.



    The miniature web graph consisting of four pages is shown in the figure above. Write the row-normalized adjacency matrix of the given web graph and apply the stochasticity and primitivity adjustments on it (clearly write the matrices being the results of each of the adjustments). All pages are initially equally important, i.e., all vertices have the same initial PageRank score. The probability of the *teleport*, i.e., the random surfer abandoning the hyperlink structure of the web graph and entering a random URL is 0.1. Assuming the PageRank scores are computed by applying the power method on the stochastically and primitively adjusted row-normalized adjacency matrix, compute the PageRank scores (for all vertices) after third iteration of the power method (i.e., compute the vector $\boldsymbol{\pi}^{(3)}$).

17. (*4 pts*) Text clustering.

    Your mixed collection contains six documents from (1) *Lord of the rings*, (2) *Game of thrones*, and (3) *Star trek*:

    - d_1: *"Frodo was carrying one ring made to rule them all"*
    - d_2: *"The darkness scared Picard as he knew the king of dragons was near"*
    - d_3: *"The king of darkness wanted his ring back from Frodo"*
    - d_4: *"Daenerys wanted her throne back and was willing to fight for it"*
    - d_5: *"Daenerys would have defeated the king, if the dragons saw in darkness"*
    - d_6: *"Looking at his ring, Picard realized he could rule but not fight."*

    The pre-built set of index terms is as follows:

    {*Frodo*, *king*, *ring*, *rule*, *fight*, *Picard*, *Daenerys*, *dragons*, *throne*, *darkness*}.

    Your task is to group the given documents into three clusters using the single-linkage agglomerative clustering. The documents are represented as vectors over given index terms and the similarity between the two documents is the cosine of the angle between the corresponding document vectors.

18. (*4 pts*) Latent semantic indexing.

Consider a word-document matrix consisting of documents $d1 \cdots d4$ and words $w1 \cdots w3$. In order to build an LSI model, an SVD of the matrix was performed as follows:

$$\begin{array}{c} \\ w_1 \\ w_2 \\ w_3 \end{array} \begin{array}{c} \begin{array}{cccc} d_1 & d_2 & d_3 & d_4 \end{array} \\ \begin{pmatrix} 5 & 3 & 0 & 1 \\ 3 & 2 & 2 & 6 \\ 0 & 0 & 8 & 7 \end{pmatrix} \end{array} = \underset{U}{\begin{pmatrix} 0.2 & 0.8 & -0.6 \\ 0.5 & 0.4 & 0.7 \\ 0.8 & -0.4 & -0.4 \end{pmatrix}} \underset{D}{\begin{pmatrix} 12.3 & 0.0 & 0.0 & 0.0 \\ 0.0 & 6.7 & 0.0 & 0.0 \\ 0.0 & 0.0 & 2.1 & 0.0 \end{pmatrix}} \underset{V^T}{\begin{pmatrix} 0.2 & 0.1 & 0.6 & 0.7 \\ 0.8 & 0.5 & -0.4 & 0.0 \\ -0.3 & -0.1 & -0.7 & 0.7 \\ 0.0 & 0.0 & 0.0 & 0.0 \end{pmatrix}}$$

(a) Compute the reconstructed version of document d_2 using only the two most important latent concepts (k=2).

(b) Which concept (u_1, u_2 or u_3) has the most impact on reconstructing the document d_3?

(c) Assuming we use only two latent concepts for reconstruction ($k = 2$), compute the similarity (any of the measures mentioned in class) of d_1 and d_2 *without* explicitly computing reconstructed documents.

19. (*4 pts*) Language modeling IR.

Consider the following collection of documents:
d_1 = *"While hobbits have hairy feet, elves do not."*
d_2 = *"Vulcans have pointy ears, elves have those ears too."*
and a query q = "*elves ears shape*".

(a) From the collection, compute all required parameters for a retrieval model based on language modeling (query retrieval model). Use unigram language models with Laplace smoothing ($\alpha = 0.5$). The vocabulary on which the models should be built and used includes the following words: {*hobbits*, *hairy*, *feet*, *elves*, *Vulcans*, *pointy*, *ears*, *shape* }. All other words may be ignored.

(b) Use the derived retrieval model to score documents d_1 and d_2 with respect to the query q.

20. (*4 pts*) Annotation and evaluation.

Two annotators are annotating part-of-speech tags for content words (only nouns, verbs, and adjectives). Their annotations for ten different tokens are given in the first two columns of the table below. The annotators resolved their disagreements by consensus, producing that way the gold standard annotations (third column in the table) for the evaluation of the supervised model. The predictions of the existing POS tagger on the same set of tokens are given in the last column of the table.

Ann. #1	Ann. #2	Gold standard	POS tagger
noun	*noun*	*noun*	*noun*
noun	*adj*	*adj*	*noun*
adj	*adj*	*adj*	*adj*
verb	*verb*	*verb*	*verb*
verb	*verb*	*verb*	*adj*
adj	*noun*	*adj*	*noun*
noun	*noun*	*noun*	*noun*
verb	*noun*	*noun*	*noun*
verb	*verb*	*verb*	*verb*
noun	*noun*	*noun*	*adj*

(a) Compute the annotation agreement (IAA) between the two annotators in terms of Cohen's kappa (κ).

(b) Briefly explain why kappa coefficient in most cases is a more realistic measure of interannotator agreement than agreement ratio.

(c) Evaluate the performance of the POS tagger in terms of micro- and macro-averaged *accuracy*, *precision*, *recall*, and F_1-score (i.e., compute Acc^{μ}, P^{μ}, R^{μ}, ${F_1}^{\mu}$, Acc^{M}, P^{M}, R^{M}, ${F_1}^{M}$)