University of Zagreb
Faculty of Electrical Engineering and Computing

September 10, 2014

Text Analysis and Retreival – Second Re-Exam (AY 2013/2014)

*The exam has **24 questions** for a total of **40 points**. Multi-choice questions carry 1 point each (1/2 point subtracted for incorrect answer), while the remaining four questions carry 5 points each. The page limit per essay question is two A4 pages. The exam duration is **150 minutes**. You must turn in the exam questions with your solutions.*

Part A: Multi-choice questions (*20 points*)

1. (*1 pt*) Consider the sentence *"Luis Suarez is [banned] from world football for four months for [biting] Italy defender Giorgio Chiellini."*. To be able to determine that *banning* occurred after *biting*, the system must be capable of doing:

   (a) event coreference resolution (c) entity disambiguation
   (b) temporal expression extraction (d) temporal relation extraction

2. (*1 pt*) You're classyfing news into topic categories using SVM. Somehow you've figured out that the first three words of the title are very good indicators of the correct category. Vocabulary of the training set consists of 10K words. How many numeric features will you need to encode the first three words of the title using one-hot encoding? (Assume that no title is shorter than three words.)

   (a) 3 features (b) 42 features (c) 10K features (d) 30K features

3. (*1 pt*) Which type of feature is prone to cause overfitting when used for authorship attribution?

   (a) Word length (b) Character n-grams (c) part-of-speech patterns (d) Content words

4. (*1 pt*) Many QA systems search for an answer by first determining the correct answer type. What type of questions are often ambiguous when it comes to determining the correct answer type?

   (a) *What* questions (b) *Where* questions (c) *Who* questions (d) *When* questions

5. (*1 pt*) You're doing text classification and wish to preserve some rudimentary syntactic information. What features will you use?

   (a) stop words (b) words (c) trigrams (d) capitalized words

6. (*1 pt*) What would be the typical sequence of tasks in an NLP pipeline consisting of part of speech tagging (POS), named entity recognition (NER), relation extraction (RE), coreference resolution (CR), entity linking (EL)?

   (a) POS → NET → EL → RE → CR (c) POS → NER → RE → CR → EL
   (b) POS → NER → CR → RE → EL (d) POS → CR → NER → RE → EL

7. (*1 pt*) For a query q, there are four documents in the collection that are relevant (R), while the rest is not relevant (N). Given q, the system returns six documents: N, R, N, R, N, N. What's the system's F1-score?

   (a) 16.7% (b) 41.7% (c) 40% (d) 50%

8. (*1 pt*) You've hired two annotators to label 1000 tweets whether they are subjective or not. The annotators agree on 900 tweets, of which 200 they've labeled as subjective. Of remaining 100 tweets, on which the annotators don't agree, only 10 were labeled as subjective by one of the annotators. What is the interannotator agreement in terms of the kappa score?

   (a) 0.91 (b) 0.622 (c) 0.736 (d) 0.9

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

   (a) Is more computationally efficient (c) Gives a relevance score which is not binary

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

10. (*1 pt*) Which of the following is the true shortcoming of the dictionary-based construction of a sentiment lexicon?

    (a) Dictionary-based approaches cannot be used in a semi-supervised setting, which is natural for constructing a sentiment lexicon (because we start from a small set of words with known sentiment orientation)

    (b) Dictionary-based approach find is unable to find the very general sentiment clues such as "good" or "bad"

    (c) Semantic relations between concepts in a dictionary are too crisp to use them to propagate sentiment

    (d) Dictionary-based approach is unable to find sentiment clues with domain-specific orientations

11. (*1 pt*) Which of the following tasks is likely to benefit from coreference resolution?

    (a) Stemming (b) Text classification (c) Parsing (d) Relation extraction

12. (*1 pt*) Distributional similarity (DS) and WordNet represent two complementary approaches to lexical semantics. What is the advantage of using DS over WordNet?

    (a) Can distinguish between different senses of a word

    (b) Cheaper to build

    (c) Offers descriptions (glosses) for each word

    (d) Can be manually edited

13. (*1 pt*) Which of the following NLP tools works at the semantic level?

    (a) parser (b) word sense disambiguation algorithm (c) stemmer (d) lemmatizer

14. (*1 pt*) In NLP, using sequence labeling models such as HMM and CRF rather than performing sequence labeling as classification is preferred because:

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

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

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

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

15. (*1 pt*) Given the question *"When did Jack London die?"*, a QA systems returns the sentence *" Following Jack London's death in November 1916, a biographical myth developed in which he has been portrayed as an alcoholic womanizer who committed suicide."* Based on this one example, we may speculate that the system is capable of:

    (a) answer fusion (c) interactive QA

    (b) factoid QA with answer generation (d) factoid QA with answer extraction

16. (*1 pt*) Supervised models for text similarity are most often evaluated using:

    (a) Correlation (b) Recall (c) Squared error (d) F-score

17. (*1 pt*) After performing inference for Latent Dirichlet Allocation, a latent document representation is:

    (a) A distribution over topics

    (b) Columns of the U matrix in SVD decomposition

    (c) A Dirichlet distribution

    (d) Columns of the V^T matrix in SVD decomposition

18. (*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) T → SS → P → L → PT (c) SS → T → P → L → PT

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

19. (*1 pt*) What is the name of the IR technique that automatically refines the original query by treating the top k initially retrieved documents as relevant?

    (a) pseudo-relevance feedback (c) query expansion

    (b) distributional similarity (d) sequence labeling

20. (*1 pt*) When framed as a machine learning problem, named entity classification is typically framed as a:

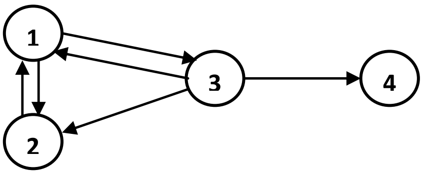
    (a) sequence labeling problem (c) binary classification problem

    (b) multi-labeling problem (d) clustering problem

Part B: Problem questions (*10 points*)

21. (*5 pts*) The PageRank algorithm.

    The miniature web graph consisting of four pages is shown in the figure below. 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 0.15. Assuming the PageRank scores are computed by applying the power method on the stochastically and primitively adjusted row-normalized adjacency matrix, which is the most relevant page (according to the PageRank scores) after two iterations of the power method?



22. (*5 pts*) Probabilistic IR models.

    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"*

    The pre-built set of index terms is as follows: {*Frodo*, *king*, *ring*, *rule*, *Daenerys*, *throne*}.

    Your task is to rank documents according to their relevance to the query *"Picard, Daenerys, Frodo, and the dragons on the throne"* using the probabilistic binary independence model given by the expression (3) (using the common practice approximations $p(D_t|Q, r) = 0.5$ and $p(D_t|Q, \bar{r}) = n_t/N_d$).

$$\sum_{t \in q} \log \frac{p(D_t|q, r)}{p(D_t|q, \bar{r})} \tag{1}$$

Part C: Essay questions (*10 points*)

23. (*5 pts*) Question answering.

    An internet security company asks you to develop a system capable of providing answers to both very specific and more generic questions regarding internet security posed by their users. The company keeps various logs for their customers and wants to automatically answer specific user questions (for which the answer can be generated from the logs) such as:

> *"When was the last time IP address 161.53.79.11 connected to my server?"*
> *"What's the top threat source that's been active over the past year?"*
> *"Which is the least trustworthy ISP in Belize?"*

Additionally, the company wants to automatically provide answers to more general questions (by searching the web), such as following:

> *"Which application typically uses port 8080?"*
> *"What is a buffer overflow attack?"*
> *"What is a DDoS attack?"*

Unlike the specific questions, the general questions typically do not have a single piece of information as answer, but rather require a more descriptive answer (at least a sentence or a paragraph). The answer provided to the user in case of general questions should be merged from several top web-search results. The answer should be coherent, should not contain redundant information and should be at most ten sentences long.

How would you discriminate the specific from general questions? How would you find sources of information for the general questions? How would you find the answers to specific questions? How would you form the answers to general questions? Discuss and elaborate each of your decisions on the system design.

24. (*5 pts*) Sentiment analysis.

    Presidential elections are closing in and the presidential candidates would like to assess their popularity based on what is being written about them by regular internet users on different social platforms (social network posts, tweets, blog posts, forum comments, etc.). An example of a comment about a politician "Conan Barbarian" is given in the following example.

> *It is only the brave moves of Conan Barbarian that got this goddamn country out of the crisis. Conan, just keep on the awesomeeeeee work! Conan Barbarian FTW!!!!!!*

Your task is to automatically analyze the user generated content from different social platforms and (1) recognize the politicians being mentioned, (2) recognize the sentiment expressed towards them, and (3) produce a final "popularity score" for each of the politicians. Elaborate on how you would solve each of these three tasks. Are there any additional preprocessing steps you would need to perform?

University of Zagreb
Faculty of Electrical Engineering and Computing

September 10, 2014

Text Analysis and Retreival – Second Re-Exam (AY 2013/2014)

The exam has ***24 questions*** *for a total of* ***40 points****. Multi-choice questions carry 1 point each (1/2 point subtracted for incorrect answer), while the remaining four questions carry 5 points each. The page limit per essay question is two A4 pages. The exam duration is* ***150 minutes****. You must turn in the exam questions with your solutions.*

Part A: Multi-choice questions (*20 points*)

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

   (a) Labels of tokens are not dependent on the labels of other tokens in the sequence
   (b) HMM and CRF allow to integrate labels from both side surrounding tokens as features
   (c) There is an assumption of independence of individual classification decisions
   (d) The uncertainty of token-wise decisions is not propagated

2. (*1 pt*) A user wants to know the answer to the question *"What did the Yalta Conference lead to?"*. The only document that contains the answer contains the passage that reads as follows: *"The Yalta Conference was held in the city of Yalta, Crimea. It was held in an atmosphere of mistrust and eventually lead to the start of the cold war"*. What should the system certainly be capable of doing, if it is to produce the answer *"The Yalta Conference lead to the start of the cold war."*?

   (a) keyword extraction (b) text classification (c) answer generation (d) simple reasoning

3. (*1 pt*) What would be the typical sequence of tasks in an NLP pipeline consisting of part of speech tagging (POS), named entity recognition (NER), relation extraction (RE), coreference resolution (CR), entity linking (EL)?

   (a) POS → RE → POS → EL → CR (c) NER → POS → RE → CR → EL
   (b) POS → NER → RE → CR → EL (d) POS → NER → CR → RE → EL

4. (*1 pt*) What is the name of the IR technique that automatically refines the original query by treating the top k initially retrieved documents as relevant?

   (a) pseudo-relevance feedback (b) faceted search (c) query expansion (d) relevance feedback

5. (*1 pt*) When framed as a machine learning problem, relation extraction is typically framed as a:

   (a) multi-labeling problem (c) sequence labeling problem
   (b) regression problem (d) multi-class classification problem

6. (*1 pt*) In a typical NLP pipeline, shallow parser (aka chunking) usually comes immediately after:

   (a) Parsing (b) Tokenization (c) POS tagging and lemmatization (d) Stemming

7. (*1 pt*) Distributional similarity (DS) and WordNet represent two complementary approaches to lexical semantics. What is the advantage of using DS over WordNet?

   (a) Can distinguish between different senses of a word
   (b) Can be manually edited
   (c) Can detect both similarity and relatedness
   (d) Offers descriptions (glosses) for each word

8. (*1 pt*) After performing inference for Latent Dirichlet Allocation, a latent document representation is:

   (a) A distribution over topics (c) A distribution over words

   (b) A Dirichlet distribution (d) Columns of the U matrix in SVD decomposition

9. (*1 pt*) For a query q, there are four documents in the collection that are relevant (R), while the rest is not relevant (N). Given q, the system returns six documents: N, R, N, R, N, N. What's the system's F1-score?

   (a) 40% (b) 50% (c) 16.7% (d) 41.7%

10. (*1 pt*) You're doing text classification and wish to preserve some rudimentary syntactic information. What features will you use?

    (a) words (b) stemmed words (c) capitalized words (d) bigrams

11. (*1 pt*) Which of the following is the true shortcoming of the dictionary-based construction of a sentiment lexicon?

    (a) Semantic relations between concepts in a dictionary are too crisp to use them to propagate sentiment

    (b) Dictionary-based approach is unable to find sentiment clues with domain-specific orientations

    (c) The label propagation algorithms applied on the graphs built within dictionary-based approaches do not always converge

    (d) Dictionary-based approaches cannot be used in a semi-supervised setting, which is natural for constructing a sentiment lexicon (because we start from a small set of words with known sentiment orientation)

12. (*1 pt*) What aspect of language is concerned with how words inflect for case, gender, and number?

    (a) Pragmatics (b) Morphology (c) Syntax (d) Phonology

13. (*1 pt*) Which of the following tasks is likely to benefit from coreference resolution?

    (a) Text classification (c) Part of speech tagging

    (b) Relation extraction (d) Document clustering

14. (*1 pt*) How does the *pyramid method* for evaluating automatic summarization work?

    (a) a correlation between numerical human ratings and system outputs is computed

    (b) humans subjectively rate the generated summaries on a numeric scale

    (c) an automated method compares the generated summaries to model (referent) summaries

    (d) humans compare the model (referent) summaries to generated summaries using a well-defined methodology

15. (*1 pt*) Which type of feature is prone to cause overfitting when used for authorship attribution?

    (a) part-of-speech patterns (b) Content words (c) Function words (d) Character n-grams

16. (*1 pt*) The *main* advantage of the two-Poisson model over the binary relevance retrieval model is that it:

    (a) Accounts for word frequencies (c) Gives a relevance score which is not binary

    (b) Accounts for word order (d) Accounts for document length

17. (*1 pt*) Many QA systems search for an answer by first determining the correct answer type. What type of questions are often ambiguous when it comes to determining the correct answer type?

    (a) *How* questions (b) *Who* questions (c) *What* questions (d) *Where* questions

18. (*1 pt*) You've hired two annotators to label 1000 tweets whether they are subjective or not. The annotators agree on 800 tweets, of which 100 they've labeled as subjective. Of remaining 200 tweets, on which the annotators don't agree, only 50 were labeled as subjective by one of the annotators. What is the interannotator agreement in terms of the kappa score?

    (a) 0.675 (b) 0.8 (c) 0.385 (d) 0.81

19. (*1 pt*) Consider the sentence *"Luis Suarez is [banned] from world football for four months for [biting] Italy defender Giorgio Chiellini."*. To be able to determine that *banning* occurred after *biting*, the system must be capable of doing:

    (a) relation extraction (c) temporal expression extraction

    (b) temporal relation extraction (d) named entity recognition

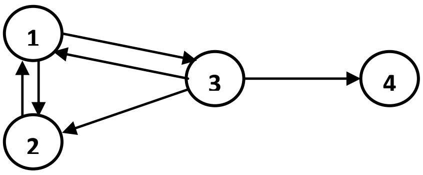
20. (*1 pt*) You're classyfing news into topic categories using SVM. Somehow you've figured out that the first three words of the title are very good indicators of the correct category. Vocabulary of the training set consists of 10K words. How many numeric features will you need to encode the first three words of the title using one-hot encoding? (Assume that no title is shorter than three words.)

    (a) 10K features (b) 30K features (c) 3 features (d) 42 features

Part B: Problem questions (*10 points*)

21. (*5 pts*) The PageRank algorithm.

    The miniature web graph consisting of four pages is shown in the figure below. 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 0.15. Assuming the PageRank scores are computed by applying the power method on the stochastically and primitively adjusted row-normalized adjacency matrix, which is the most relevant page (according to the PageRank scores) after two iterations of the power method?



22. (*5 pts*) Probabilistic IR models.

    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"*

    The pre-built set of index terms is as follows: {*Frodo*, *king*, *ring*, *rule*, *Daenerys*, *throne*}.

    Your task is to rank documents according to their relevance to the query *"Picard, Daenerys, Frodo, and the dragons on the throne"* using the probabilistic binary independence model given by the expression (3) (using the common practice approximations $p(D_t|Q, r) = 0.5$ and $p(D_t|Q, \bar{r}) = n_t/N_d$).

$$\sum_{t \in q} \log \frac{p(D_t|q, r)}{p(D_t|q, \bar{r})} \tag{2}$$

Part C: Essay questions (*10 points*)

23. (*5 pts*) Question answering.

    An internet security company asks you to develop a system capable of providing answers to both very specific and more generic questions regarding internet security posed by their users. The company keeps various logs for their customers and wants to automatically answer specific user questions (for which the answer can be generated from the logs) such as:

> *"When was the last time IP address 161.53.79.11 connected to my server?"*
> *"What's the top threat source that's been active over the past year?"*
> *"Which is the least trustworthy ISP in Belize?"*

Additionally, the company wants to automatically provide answers to more general questions (by searching the web), such as following:

> *"Which application typically uses port 8080?"*
> *"What is a buffer overflow attack?"*
> *"What is a DDoS attack?"*

Unlike the specific questions, the general questions typically do not have a single piece of information as answer, but rather require a more descriptive answer (at least a sentence or a paragraph). The answer provided to the user in case of general questions should be merged from several top web-search results. The answer should be coherent, should not contain redundant information and should be at most ten sentences long.

How would you discriminate the specific from general questions? How would you find sources of information for the general questions? How would you find the answers to specific questions? How would you form the answers to general questions? Discuss and elaborate each of your decisions on the system design.

24. (*5 pts*) Sentiment analysis.

    Presidential elections are closing in and the presidential candidates would like to assess their popularity based on what is being written about them by regular internet users on different social platforms (social network posts, tweets, blog posts, forum comments, etc.). An example of a comment about a politician "Conan Barbarian" is given in the following example.

> *It is only the brave moves of Conan Barbarian that got this goddamn country out of the crisis. Conan, just keep on the awesomeeeeee work! Conan Barbarian FTW!!!!!!*

Your task is to automatically analyze the user generated content from different social platforms and (1) recognize the politicians being mentioned, (2) recognize the sentiment expressed towards them, and (3) produce a final "popularity score" for each of the politicians. Elaborate on how you would solve each of these three tasks. Are there any additional preprocessing steps you would need to perform?

University of Zagreb
Faculty of Electrical Engineering and Computing

September 10, 2014

Text Analysis and Retreival – Second Re-Exam (AY 2013/2014)

The exam has ***24 questions*** *for a total of* ***40 points****. Multi-choice questions carry 1 point each (1/2 point subtracted for incorrect answer), while the remaining four questions carry 5 points each. The page limit per essay question is two A4 pages. The exam duration is* ***150 minutes****. You must turn in the exam questions with your solutions.*

Part A: Multi-choice questions (*20 points*)

1. (*1 pt*) For a query q, there are four documents in the collection that are relevant (R), while the rest is not relevant (N). Given q, the system returns six documents: N, R, N, R, N, N. What's the system's F1-score?

   (a) 40% (b) 41.7% (c) 33% (d) 16.7%

2. (*1 pt*) What would be the typical sequence of tasks in an NLP pipeline consisting of part of speech tagging (POS), named entity recognition (NER), relation extraction (RE), coreference resolution (CR), entity linking (EL)?

   (a) NER → POS → RE → CR → EL (c) POS → RE → POS → EL → CR

   (b) POS → CR → NER → RE → EL (d) POS → NER → CR → RE → EL

3. (*1 pt*) Consider the sentence *"Luis Suarez is [banned] from world football for four months for [biting] Italy defender Giorgio Chiellini."*. To be able to determine that *banning* occurred after *biting*, the system must be capable of doing:

   (a) relation extraction (c) entity disambiguation

   (b) temporal expression extraction (d) event extraction

4. (*1 pt*) After performing inference for Latent Dirichlet Allocation, a latent document representation is:

   (a) A distribution over topics (c) Columns of the U matrix in SVD decomposition

   (b) A distribution over words (d) A Dirichlet distribution

5. (*1 pt*) Distributional similarity (DS) and WordNet represent two complementary approaches to lexical semantics. What is the advantage of using DS over WordNet?

   (a) Offers descriptions (glosses) for each word

   (b) Can be manually edited

   (c) Can be easily adapted to different domains

   (d) Can distinguish between different senses of a word

6. (*1 pt*) A user wants to know the answer to the question *"What did the Yalta Conference lead to?"*. The only document that contains the answer contains the passage that reads as follows: *"The Yalta Conference was held in the city of Yalta, Crimea. It was held in an atmosphere of mistrust and eventually lead to the start of the cold war"*. What should the system certainly be capable of doing, if it is to produce the answer *"The Yalta Conference lead to the start of the cold war."*?

   (a) answer generation (b) simple reasoning (c) answer fusion (d) answer extraction

7. (*1 pt*) The *main* advantage of the two-Poisson model over the binary relevance retrieval model is that it:

   (a) Accounts for word frequencies (c) Accounts for document length

   (b) Is more computationally efficient (d) Gives a relevance score which is not binary

8. (*1 pt*) Which of the following tasks is likely to benefit from coreference resolution?

   (a) Relation extraction (b) Stemming (c) Parsing (d) Part of speech tagging

9. (*1 pt*) When framed as a machine learning problem, named entity classification is typically framed as a:

   (a) sequence labeling problem (c) regression problem

   (b) multi-labeling problem (d) clustering problem

10. (*1 pt*) Which of the following is the true shortcoming of the dictionary-based construction of a sentiment lexicon?

    (a) Dictionary-based approach is unable to find sentiment clues with domain-specific orientations

    (b) Dictionary-based approaches cannot be used in a semi-supervised setting, which is natural for constructing a sentiment lexicon (because we start from a small set of words with known sentiment orientation)

    (c) The label propagation algorithms applied on the graphs built within dictionary-based approaches do not always converge

    (d) Dictionary-based approach find is unable to find the very general sentiment clues such as “good” or “bad”

11. (*1 pt*) You’re classyfing news into topic categories using SVM. Somehow you’ve figured out that the first three words of the title are very good indicators of the correct category. Vocabulary of the training set consists of 10K words. How many numeric features will you need to encode the first three words of the title using one-hot encoding? (Assume that no title is shorter than three words.)

    (a) 42 features (b) 3 features (c) 30K features (d) 10K features

12. (*1 pt*) You’re doing text classification and wish to preserve some rudimentary syntactic information. What features will you use?

    (a) stemmed words (b) stop words (c) trigrams (d) capitalized words

13. (*1 pt*) You’ve hired two annotators to label 1000 tweets whether they are subjective or not. The annotators agree on 800 tweets, of which 100 they’ve labeled as subjective. Of remaining 200 tweets, on which the annotators don’t agree, only 50 were labeled as subjective by one of the annotators. What is the interannotator agreement in terms of the kappa score?

    (a) 0.385 (b) 0.675 (c) 0.81 (d) 0.736

14. (*1 pt*) Which type of feature is prone to cause overfitting when used for authorship attribution?

    (a) Word length (b) Character n-grams (c) Content words (d) Function words

15. (*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 → L → PT → P (c) SS → T → P → L → PT

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

16. (*1 pt*) Which of the following NLP tools works at the semantic level?

    (a) word sense disambiguation algorithm (b) parser (c) POS tagger (d) lemmatizer

17. (*1 pt*) What is the name of the IR technique that automatically refines the original query by treating the top k initially retrieved documents as relevant?

    (a) pseudo-relevance feedback (c) faceted search

    (b) query expansion (d) distributional similarity

18. (*1 pt*) Supervised models for text similarity are most often evaluated using:

    (a) ROUGE score (b) Precision (c) Correlation (d) Accuracy

19. (*1 pt*) Many QA systems search for an answer by first determining the correct answer type. What type of questions are often ambiguous when it comes to determining the correct answer type?

    (a) *Who* questions (b) *What* questions (c) *How* questions (d) *Where* questions

20. (*1 pt*) In NLP, using sequence labeling models such as HMM and CRF rather than performing sequence labeling as classification is preferred because:

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

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

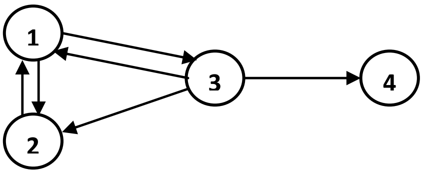
    (c) Any classification algorithm (e.g., naïve Bayes, SVM) can be pluged into a sequence labeling models such as HMM and CRF

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

Part B: Problem questions (*10 points*)

21. (*5 pts*) The PageRank algorithm.

    The miniature web graph consisting of four pages is shown in the figure below. 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 0.15. Assuming the PageRank scores are computed by applying the power method on the stochastically and primitively adjusted row-normalized adjacency matrix, which is the most relevant page (according to the PageRank scores) after two iterations of the power method?



22. (*5 pts*) Probabilistic IR models.

    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"*

    The pre-built set of index terms is as follows: {*Frodo*, *king*, *ring*, *rule*, *Daenerys*, *throne*}.

    Your task is to rank documents according to their relevance to the query *"Picard, Daenerys, Frodo, and the dragons on the throne"* using the probabilistic binary independence model given by the expression (3) (using the common practice approximations $p(D_t|Q,r) = 0.5$ and $p(D_t|Q,\bar{r}) = n_t/N_d$).

$$\sum_{t \in q} \log \frac{p(D_t|q,r)}{p(D_t|q,\bar{r})} \tag{3}$$

Part C: Essay questions (*10 points*)

23. (*5 pts*) Question answering.

    An internet security company asks you to develop a system capable of providing answers to both very specific and more generic questions regarding internet security posed by their users. The company keeps various logs for their customers and wants to automatically answer specific user questions (for which the answer can be generated from the logs) such as:

> *"When was the last time IP address 161.53.79.11 connected to my server?"*
> *"What's the top threat source that's been active over the past year?"*
> *"Which is the least trustworthy ISP in Belize?"*

    Additionally, the company wants to automatically provide answers to more general questions (by searching the web), such as following:

> *"Which application typically uses port 8080?"*
> *"What is a buffer overflow attack?"*
> *"What is a DDoS attack?"*

    Unlike the specific questions, the general questions typically do not have a single piece of information as answer, but rather require a more descriptive answer (at least a sentence or a paragraph). The answer provided to the user in case of general questions should be merged from several top web-search results. The answer should be coherent, should not contain redundant information and should be at most ten sentences long.

    How would you discriminate the specific from general questions? How would you find sources of information for the general questions? How would you find the answers to specific questions? How would you form the answers to general questions? Discuss and elaborate each of your decisions on the system design.

24. (*5 pts*) Sentiment analysis.

    Presidential elections are closing in and the presidential candidates would like to assess their popularity based on what is being written about them by regular internet users on different social platforms (social network posts, tweets, blog posts, forum comments, etc.). An example of a comment about a politician "Conan Barbarian" is given in the following example.

> *It is only the brave moves of Conan Barbarian that got this goddamn country out of the crisis. Conan, just keep on the awesomeeeeee work! Conan Barbarian FTW!!!!!!*

    Your task is to automatically analyze the user generated content from different social platforms and (1) recognize the politicians being mentioned, (2) recognize the sentiment expressed towards them, and (3) produce a final "popularity score" for each of the politicians. Elaborate on how you would solve each of these three tasks. Are there any additional preprocessing steps you would need to perform?

University of Zagreb
Faculty of Electrical Engineering and Computing

September 10, 2014

Text Analysis and Retreival – Second Re-Exam (AY 2013/2014)

*The exam has **24 questions** for a total of **40 points**. Multi-choice questions carry 1 point each (1/2 point subtracted for incorrect answer), while the remaining four questions carry 5 points each. The page limit per essay question is two A4 pages. The exam duration is **150 minutes**. You must turn in the exam questions with your solutions.*

Part A: Multi-choice questions (*20 points*)

1. (*1 pt*) After performing inference for Latent Dirichlet Allocation, a latent document representation is:

   (a) A Dirichlet distribution (c) Columns of the U matrix in SVD decomposition
   (b) A distribution over words (d) A distribution over topics

2. (*1 pt*) Consider the sentence *"Luis Suarez is [banned] from world football for four months for [biting] Italy defender Giorgio Chiellini."*. To be able to determine that *banning* occurred after *biting*, the system must be capable of doing:

   (a) named entity recognition (c) event extraction
   (b) event coreference resolution (d) entity disambiguation

3. (*1 pt*) The *main* advantage of the BM11 and BM25 algorithms, compared to the two-Poisson model is that they:

   (a) Account for document length (c) Account for word frequencies
   (b) Account for word order (d) Consider word dependencies within a document

4. (*1 pt*) What would be the typical sequence of tasks in an NLP pipeline consisting of part of speech tagging (POS), named entity recognition (NER), relation extraction (RE), coreference resolution (CR), entity linking (EL)?

   (a) POS → RE → POS → EL → CR (c) POS → NER → CR → RE → EL
   (b) POS → NET → EL → RE → CR (d) NER → POS → RE → CR → EL

5. (*1 pt*) Which type of feature is prone to cause overfitting when used for authorship attribution?

   (a) Word length (b) Function words (c) Character n-grams (d) Content words

6. (*1 pt*) Which of the following is the true shortcoming of the dictionary-based construction of a sentiment lexicon?

   (a) Semantic relations between concepts in a dictionary are too crisp to use them to propagate sentiment
   (b) Dictionary-based approach find is unable to find the very general sentiment clues such as "good" or "bad"
   (c) Dictionary-based approach is unable to find sentiment clues with domain-specific orientations
   (d) The label propagation algorithms applied on the graphs built within dictionary-based approaches do not always converge

7. (*1 pt*) Which of the following tasks is likely to benefit from coreference resolution?

   (a) Relation extraction (b) Text classification (c) Part of speech tagging (d) Stemming

8. (*1 pt*) When framed as a machine learning problem, relation extraction is typically framed as a:

   (a) multi-class classification problem (c) binary classification problem

   (b) sequence labeling problem (d) clustering problem

9. (*1 pt*) How does the *pyramid method* for evaluating automatic summarization work?

   (a) a correlation between numerical human ratings and system outputs is computed

   (b) an automated method compares the generated summaries to model (referent) summaries

   (c) an automated method compares the generated summaries to the original text

   (d) humans compare the model (referent) summaries to generated summaries using a well-defined methodology

10. (*1 pt*) You've hired two annotators to label 1000 tweets whether they are subjective or not. The annotators agree on 900 tweets, of which 200 they've labeled as subjective. Of remaining 100 tweets, on which the annotators don't agree, only 10 were labeled as subjective by one of the annotators. What is the interannotator agreement in terms of the kappa score?

    (a) 0.9 (b) 0.736 (c) 0.675 (d) 0.91

11. (*1 pt*) Distributional similarity (DS) and WordNet represent two complementary approaches to lexical semantics. What is the advantage of using WordNet over DS?

    (a) Gives a graded notion of semantic similarity

    (b) Can detect both similarity and relatedness

    (c) Offers descriptions (glosses) for each word

    (d) Cheaper to build

12. (*1 pt*) What is the name of the IR technique that automatically refines the original query by treating the top k initially retrieved documents as relevant?

    (a) query expansion (c) pseudo-relevance feedback

    (b) sequence labeling (d) distributional similarity

13. (*1 pt*) You're classyfing news into topic categories using SVM. Somehow you've figured out that the first three words of the title are very good indicators of the correct category. Vocabulary of the training set consists of 10K words. How many numeric features will you need to encode the first three words of the title using one-hot encoding? (Assume that no title is shorter than three words.)

    (a) 3 features (b) 30K features (c) 15 features (d) 42 features

14. (*1 pt*) Many QA systems search for an answer by first determining the correct answer type. What type of questions are often ambiguous when it comes to determining the correct answer type?

    (a) *Who* questions (b) *When* questions (c) *What* questions (d) *How* questions

15. (*1 pt*) A user wants to know the answer to the question *"What did the Yalta Conference lead to?"*. The only document that contains the answer contains the passage that reads as follows: *"The Yalta Conference was held in the city of Yalta, Crimea. It was held in an atmosphere of mistrust and eventually lead to the start of the cold war"*. What should the system certainly be capable of doing, if it is to produce the answer *"The Yalta Conference lead to the start of the cold war."*?

    (a) coreference resolution (b) answer extraction (c) simple reasoning (d) keyword extraction

16. (*1 pt*) You're doing text classification and wish to preserve some rudimentary syntactic information. What features will you use?

    (a) trigrams (b) stop words (c) stemmed words (d) capitalized words

17. (*1 pt*) Prior to parsing a sentence, the sentence should be:

    (a) Both lemmatized and POS tagged (c) Lemmatized and chunked

    (b) Lemmatized but not POS tagged (d) POS tagged but not lemmatized

18. (*1 pt*) In NLP, using sequence labeling models such as HMM and CRF rather than performing sequence labeling as classification is preferred because:

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

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

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

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

19. (*1 pt*) For a query q, there are four documents in the collection that are relevant (R), while the rest is not relevant (N). Given q, the system returns six documents: N, R, N, R, N, N. What's the system's F1-score?

    (a) 33% (b) 41.7% (c) 16.7% (d) 40%

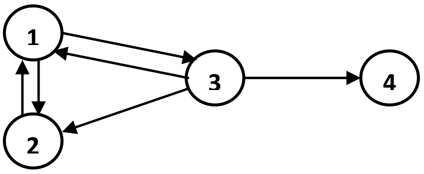
20. (*1 pt*) Which of the following NLP tools works at the semantic level?

    (a) stemmer (b) parser (c) lemmatizer (d) word sense disambiguation algorithm

Part B: Problem questions (*10 points*)

21. (*5 pts*) The PageRank algorithm.

    The miniature web graph consisting of four pages is shown in the figure below. 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 0.15. Assuming the PageRank scores are computed by applying the power method on the stochastically and primitively adjusted row-normalized adjacency matrix, which is the most relevant page (according to the PageRank scores) after two iterations of the power method?



22. (*5 pts*) Probabilistic IR models.

    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"*

    The pre-built set of index terms is as follows: {*Frodo*, *king*, *ring*, *rule*, *Daenerys*, *throne*}.

    Your task is to rank documents according to their relevance to the query *"Picard, Daenerys, Frodo, and the dragons on the throne"* using the probabilistic binary independence model given by the expression (3) (using the common practice approximations $p(D_t|Q,r) = 0.5$ and $p(D_t|Q,\bar{r}) = n_t/N_d$).

$$\sum_{t \in q} \log \frac{p(D_t|q,r)}{p(D_t|q,\bar{r})} \tag{4}$$

Part C: Essay questions (*10 points*)

23. (*5 pts*) Question answering.

    An internet security company asks you to develop a system capable of providing answers to both very specific and more generic questions regarding internet security posed by their users. The company keeps various logs for their customers and wants to automatically answer specific user questions (for which the answer can be generated from the logs) such as:

    > *"When was the last time IP address 161.53.79.11 connected to my server?"*
    > *"What's the top threat source that's been active over the past year?"*
    > *"Which is the least trustworthy ISP in Belize?"*

    Additionally, the company wants to automatically provide answers to more general questions (by searching the web), such as following:

    > *"Which application typically uses port 8080?"*
    > *"What is a buffer overflow attack?"*
    > *"What is a DDoS attack?"*

    Unlike the specific questions, the general questions typically do not have a single piece of information as answer, but rather require a more descriptive answer (at least a sentence or a paragraph). The answer provided to the user in case of general questions should be merged from several top web-search results. The answer should be coherent, should not contain redundant information and should be at most ten sentences long.

    How would you discriminate the specific from general questions? How would you find sources of information for the general questions? How would you find the answers to specific questions? How would you form the answers to general questions? Discuss and elaborate each of your decisions on the system design.

24. (*5 pts*) Sentiment analysis.

    Presidential elections are closing in and the presidential candidates would like to assess their popularity based on what is being written about them by regular internet users on different social platforms (social network posts, tweets, blog posts, forum comments, etc.). An example of a comment about a politician "Conan Barbarian" is given in the following example.

    > *It is only the brave moves of Conan Barbarian that got this goddamn country out of the crisis. Conan, just keep on the awesomeeeeee work! Conan Barbarian FTW!!!!!!*

    Your task is to automatically analyze the user generated content from different social platforms and (1) recognize the politicians being mentioned, (2) recognize the sentiment expressed towards them, and (3) produce a final "popularity score" for each of the politicians. Elaborate on how you would solve each of these three tasks. Are there any additional preprocessing steps you would need to perform?