EIDGENÖSSISCHE TECHNISCHE HOCHSCHULE LAUSANNE
POLITECNICO FEDERALE DI LOSANNA
SWISS FEDERAL INSTITUTE OF TECHNOLOGY LAUSANNE

SCHOOL OF COMPUTER AND COMMUNICATION SCIENCES
Introduction to Natural Language Processing (Ms)

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

INTRODUCTION TO NATURAL LANGUAGE PROCESSING
Fall 2018, Graded Quiz #4

NAME: **SCIPER:**

Instructions:

You have 45 minutes for this exam, which consists of 10 independent questions with different weights for a total of 30 points. For each question, the corresponding amount of points is indicated.

All documents are allowed, but no electronic devices (computer, cell phone, calculator, etc.).

Answer directly on the exam sheets (we expect only the final required answer, no justification). Do not provide any other paper; it will not be considered.

Introduction to Natural Language Processing
Graded Quiz #2, 2018

NAME: **SCIPER:**

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Question 1: [1 pt]

1.1 [62%, 1 pt] When a user submits a query Q to a vector space Information Retrieval (IR) system operating on a collection of documents, the main goal of the system is:
(Select only one answer; a penalty will be applied for wrong answers selected)

[84%] to summarize the documents relevant for the query Q
[91%] to filter out all the documents that strictly match the query Q
[63%] to identify the documents that are talking about the topic expressed in the query Q
[86%] to extract the correct answers to the query Q

Question 2: [3 pts]

The preprocessing steps implemented in a given IR system have produced the following set of tokens for the document "recycling aluminum can be crucial for the environment":

aluminum|Noun
crucial|Adj
environment|Noun
recycle|Verb

2.1 [42%, 2 pts] Which of the following NLP tools/resources are required to explain the presence of the token "recycle|Verb" in the produced set?
Only select the simplest one(s) that allow to produce the correct output under the assumption that only "aluminum", "be", and "environment" are non-ambiguous at the morpho-syntactic level.
(Several answers are possible; a penalty will be applied for wrong answers selected)

[88%] a stop word list
[62%] a morphological analyzer
[57%] a lexicon associating roots and PoS to surface forms

Introduction to Natural Language Processing
Graded Quiz #2, 2018

NAME: **SCIPER:**

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

[66%] a part-of-speech tagger

[83%] a probabilistic parser

NAME: **SCIPER:**

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

2.2 [-12%, 1 pt] Which of the following NLP tools/resources are required to explain the absence of any token related to the word "can" in the produced set?
Only select the simplest one(s) that allow to produce the correct output under the assumption that only "aluminum", "be", and "environment" are non-ambiguous at the morpho-syntactic level.
(Several answers are possible; a penalty will be applied for wrong answers selected)

[17%] a stop word list
[90%] a morphological analyzer
[5%] a lexicon associating roots and PoS to surface forms
[20%] a part-of-speech tagger
[91%] a probabilistic parser

Question 3: [6 pts]

3.1 [70%, 2 pts] As part of the analysis of a document collection to be used by an IR system, the occurrences of the 40'000 distinct words appearing in the collection have been counted, and the words have been ranked by decreasing number of occurrences. Knowing that the most frequent word appears 6'400 times in the document collection, what is the approximate number N of occurrences of the 100th most frequent word under the assumption of the Zipf law?
(Provide the answer in the form of an integer)

[70%] N = 64

3.2 [51%, 4 pts] A more detailed analysis shows that 50% of the distinct words appearing in the document collection correspond to grammatical words, and that, among the other ones, 1% occur at least 50 times, and 60% at most 3 times.
What is the dimensionality D of the vector space to be used for representing the document collection, under the assumption that only non-grammatical words can be considered as indexing terms, and that Luhn's principle with upper cut-off R(50) and lower cut-off r(3) is used for term filtering, where R(k) is the worst of the ranks of the

words occurring at least k times, and r(k) is the best of the ranks of the words appearing at most k times?
(Provide the answer in the form of an integer)

[51%] D = 7'800

Question 4: [1 pt]

4.1 [84%, 1 pt] In the standard tf.idf weighting scheme, what guarantees that the indexing terms that occur most often in a document are given a higher weight?
(Select only one answer; a penalty will be applied for wrong answers selected)

[84%] the tf component
[94%] the idf component
[92%] both components

Question 5: [2 pts]

5.1 [55%, 2 pts] In the framework of the IR system described in Question 3, is it possible for the document

D = "recycling aluminum can be crucial for the environment"

to be considered as relevant for the query

Q = "ecologic impact of metal"?

(Select only one answer; a penalty will be applied for wrong answers selected)

[92%] yes
[55%] no
[65%] undecidable

Question 6: [2 pts]

NAME: **SCIPER:**

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

6.1 [86%, 2 pts] What is the similarity between a document D and the document resulting from the concatenation of 10 copies of D, provided that the cosine similarity measure is used?
(Select only one answer; a penalty will be applied for wrong answers selected)

[100%] 0
[97%] 1/10
[86%] 1
[100%] 10
[90%] cannot be computed in this general case

NAME: **SCIPER:**

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Question 7: [2 pts]

7.1 [84%, 2 pts] For a given query Q, an IR system retrieves 50 documents with a precision P = 0.6. If one assumes that the total number of relevant documents for Q is N = 100, what is the recall R of the system for Q?
(Provide the answer in a form of a fraction)

[84%] R = 0.3

Question 8: [5 pts]

Two IR systems, S_1 and S_2, have been evaluated, and, for each of them, the evaluation resulted in the following (Recall, Precision) pairs:

S_1	S_2
(0.01, 0.80)	(0.01, 0.50)
(0.20, 0.40)	(0.20, 0.40)
(0.60, 0.10)	(0.60, 0.30)

8.1 [24%, 3 pts] Based on these evaluation results, which system should be selected for an IR application where it is important to find all the documents that are relevant for a given query?
(Select only one answer; a penalty will be applied for wrong answers selected)

[86%] system S_1
[24%] system S_2
[40%] cannot be decided

8.2 [63%, 2 pts] Based on these evaluation results, which system should be selected for an IR application performing general purpose IR from the Web?
(Select only one answer; a penalty will be applied for wrong answers selected)

[63%] system S_1
[83%] system S_2

NAME: **SCIPER:**

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

[81%] cannot be decided

NAME: **SCIPER:**

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Question 9: [5 pts]

For an application that needs to cluster a large document collection into topical clusters, the following approach is used:

(1) Each topic t is defined by a well-chosen query Q(t);
(2) For each query Q(t), an IR system is used to retrieve the set R(t) consisting of all the documents relevant for Q(t);
(3) The vector space model used by the IR system is exploited to compute, for each set R(t), the mean M(t) of the vectors representing the documents it contains;
(4) Each of the documents is allocated to the closest vector M(t) according to the Euclidian distance;
(5) M(t) is updated as the mean of the vectors it has been allocated to;
(6) (4) and (5) are iterated until no re-allocation is needed in (4).

9.1 [85%, 3 pts] Can it be guaranteed that the proposed method converges (i.e. stops after a finite number of iterations)?
(Select only one answer; a penalty will be applied for wrong answers selected)

[85%] yes
[92%] no
[98%] undecidable

9.2 [44%, 2 pts] How should the proposed method be considered?
(Select only one answer; a penalty will be applied for wrong answers selected)

[98%] as a supervised classification method
[47%] as an unsupervised classification method
[44%] as a hybrid classification method

NAME: **SCIPER:**

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Question 10: [3 pts]

The Naïve Bayes algorithm is used in the framework of a sentiment analysis application to determine, for any input tweet, which, among a predefined set of sentiments, best corresponds to the mood expressed in the tweet.

10.1 [88%, 1 pt] Does the performed tweet classification task have to be supervised in this case?
(Select only one answer; a penalty will be applied for wrong answers selected)

[88%] yes
[93%] no
[97%] it depends on the implementation

10.2 [65%, 2 pts] Let us assume that only two sentiments are considered ("joyful" and "sad") and that typically 70% of the tweets are "joyful".
To which sentiment would the Naïve Bayes algorithm associate a tweet indexed by only two terms w_1 and w_2, if:
• 10% of the occurrences of indexing terms in "joyful" tweets and 20% of the occurrences of indexing terms in "sad" tweets are w_1; while
• 30% of the occurrences of indexing terms in "joyful" tweets and 10% of the occurrences of indexing terms in "sad" tweets are w_2?
(Select only one answer; a penalty will be applied for wrong answers selected)

[65%] joyful
[72%] sad
[97%] undecidable