UNIVERSITY OF BIRMINGHAM

Revision Lecture 2024 - Weeks 1-4 + 10

Phil Smith

Overview

- Preparation Hints
- A Revision List
- The Exam

UNIVERSITY OF BIRMINGHAM

The Exam

- 120 minute exam
- 3 questions - 20 marks each
- Bring a calculator not capable of storing alphabetical information
- A “typical“ question will be:
- Some bookwork
- Some application of a technique to some data
- Some discussion of how it could be extended
- One question is more open-ended but the above is still relevant

how can it be applied to general problems.

Learning Outcomes

- Demonstrate an understanding of the major topics in Natural Language Processing
- Understand the role of machine learning techniques in widening the coverage of NLP systems
- Demonstrate an ability to apply knowledge-based and statistical techniques to real-world NLP problems

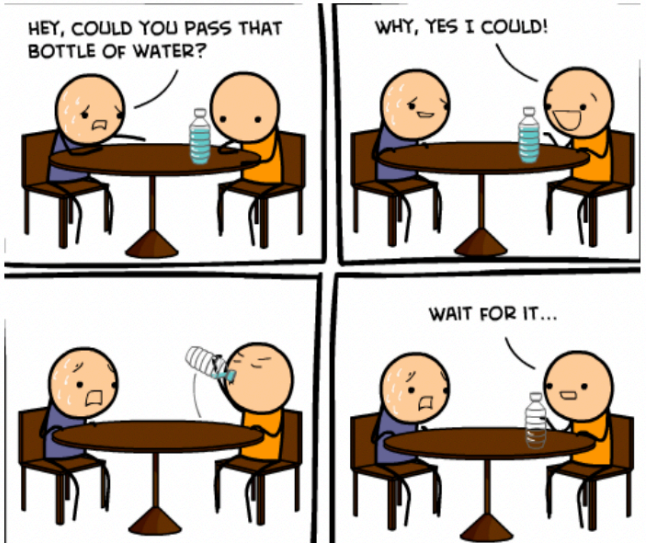
What we're looking for

- The right answer
  - Definitions
  - Applications of an algorithm to a simple example
- Otherwise:
- Sensible proposals:
  - Based on sensible applicable techniques
  - What data?
  - Realistic in terms of time, scalability,
  - Rough notions of accuracy
  - How do you evaluate?

UNIVERSITY OF BIRMINGHAM

The Interview Metaphor

- Answer the questions as if you were in an interview:
- Maxim of Quality
  - Do not say what you believe to be false
  - Do not say that for which you lack adequate evidence
- Maxim of Quantity
  - Make your contribution as informative as required
  - Do not make your contribution more informative than require
- Maxim of Relation
  - Be relevant
- Maxim of Manner
  - Be brief
  - Avoid ambiguity
  - Be orderly



Text Processing

- Regular Expressions
  - Formal language for specifying text strings
- Words and Corpora
  - Primary data source of NLP
- Tokenization
  - Segmenting words: space-based, byte-pair encoding algorithm
- Normalization
  - Transforming words into a standard format e.g stemming
- Minimum Edit Distance
  - Computing the similarity of two strings

UNIVERSITY OF BIRMINGHAM

Example Question (10 marks)

Given the text string below:

"Email addresses like test.email+regex@gmail.com and simple@example.com are common in data files. Make sure your regex also matches uncommon domains such as john.doe@company.co.uk."

Write a regular expression that matches all the email addresses mentioned in the text. Explain each part of your regular expression and why it is necessary. Your answer should include:

- A regular expression pattern that accurately identifies all the email addresses in the provided text. (4 marks)
- A detailed explanation of each component of your regular expression, indicating what it matches within the email format. (4 marks)
- Discuss briefly how your regular expression handles edge cases such as unusual domain extensions and the use of characters like "+" in the email username. (2 marks)

N-gram Language Modelling

- Probabilistic Language Models
  - Computing the probability of a sentence or a sequence of words. N-gram modelling
- Model Evaluation
  - Extrinisic and intrinsic, perplexity, the Shannon Game
- Zeros/Smoothing/Interpolation/Backoff
  - Zeros make it impossible to compute perplexity.
  - Redistribute the probability density to account for this
  - Backoff and interpolation slightly more sophisticated, conditioning on less context

Text Classification

- Naive Bayes Classifiers
  - Bag of words, Bayes Rule, Assumptions, Learner, Multinomial Version, Relationship to LMs
- Sentiment Classification
  - Negation, lexicons
- Evaluation
  - Precision, Recall, F1, Confusion Matrices, Significance Testing, Bootstrapping, Biases

UNIVERSITY OF BIRMINGHAM

Example Question

What are the two assumptions made when using a multinomial Naive Bayes classifier for text classification? (Give the mathematical formulation in your answer).

Word frequency and word occurrence play important roles in text classification. For sentiment analysis, discuss and argue why one approach would be preferable to the other and how this would be implemented in a Naive Bayes classification system.

Discuss how a lexicon might be used for sentiment classification.

Vector Semantics

- Lexical Semantics: Word Meanings, Lemmas and Senses
- Synonymy and Similarity
    - Semantic fields, connotation
- Vector Semantics
    - Meaning as a point in multidimensional space
    - Cosine similarity
    - Embeddings
- TF-IDF
- (P)PMI
    - Pointwise mutual information
- Word2Vec
    - Predict rather than count
    - Self-supervision
    - Skip-gram training

UNIVERSITY OF BIRMINGHAM

Transformers in Action

- Applications of Transformers
  - Encoder Models: Classification, Named Entity Recognition, Extractive QA
  - Decoder Models: Text generation
  - Seq2Seq: Summarization, Translation, Generative QA
- Transformer Implementation
  - Pipelines: tokeniser, model, post processing.
- Finetuning a Pretrained Model
  - Dynamic padding, attention masks

UNIVERSITY OF BIRMINGHAM

Example Question

Give an overview of the transformer pipeline and how each component works.

Good luck!

There are many ways to say "good luck" to students, depending on the context and the relationship you have with them. Here are a few suggestions:

1. "Good luck on your exam/test/presentation/project, I'm rooting for you!"
2. "Wishing you the best of luck in your studies and all your future endeavors."
3. "May you succeed in all your academic pursuits, and enjoy the journey along the way."
4. "I believe in you and your abilities, go out there and show them what you're made of."
5. "Remember that hard work and determination will pay off in the end. Good luck!"
6. "You've got this! Don't forget to take breaks and take care of yourself along the way."
7. "Sending positive vibes your way for a successful semester/quarter/year."
8. "May luck and success be on your side, always."