inzva Applied AI Program
Week 3

Natural Language Processing

Guide:
Ahmet Melek

Thank you for the help:
Şükrü Bezen: Word2Vec Notebook
Uras Mutlu: NLP Tasks Notebook

07.08.2021

Today's Contents

1.	**Concepts on NLP (with problem formulations)**	**10.00 - 10.20**
2.	**Problems in NLP Notebook**	**10.20 - 10.50**
	Break	10.50 - 11.00
3.	**Language Modelling with Word2Vec Notebook**	**11.00 - 11.50**
	Break	11.50 - 12.00
4.	**Text Classification Notebook**	**12.00 - 12.30**
	Lunch Break	12.30 - 13.10
5.	**Neural Machine Translation with Attention Notebook**	**13.10 - 14.20**
	Break	14.20 - 14.45
6.	**Named Entity Recognition Notebook**	**14.45 - 15.30**
	Break	15.30 - 15.50
7.	**Deploying a QA model**	**15.50 - 16.30**
8.	**Homework and Questions**	**16.30 - 17.00**
	Estimated Finish	**17.00**

Concepts on NLP

1. **Preprocessing**
   a. Corpus, Tokenizers, Vocabulary, Special Tokens, Lowercasing, Stopword and Punctuation Cleaning, Stemming and Lemmatization
2. **Obtaining Word Features**
   a. Vector Space Models (Embedding Layers), N-grams, Word Counts
3. **Models and Algorithms**
   a. Transformers (and attention), Sequential Models (RNN, GRU, LSTM), Cosine Similarity, Edit Distance
   b. Logistic Regression, Bayesian Models, kNN algorithm, Hash Tables, Index Search Algorithms, Hidden Markov Models, Viterbi Algorithm, Dynamic Programming
4. **Problems**
   a. Question Answering, Text Retrieval (Search), Text Classification, Entity Recognition, Machine Translation, Speech, Language Modelling, Text Generation
5. **Visualization**
   a. PCA, tSNE
6. **Evaluation**
   a. Recall, Precision, f1 score, Confusion Matrix, Perplexity, BLEU, ROGUE
7. **Experimentation and Development**
   a. Problem Formulation, Annotation (Dataset Creation), Preprocessing, Training, Evaluation, Experimentation, Scheduling Model Development, Inference (prediction time), Application

Problems in NLP Notebook

1. **Question Answering:**
   Can be extractive or abstractive.
   a. Extractive QA is finding the answer in the context.
   b. Abstractive QA is reading the context, then generating an answer based on the context.

   Can be context based or open domain.

   a. Context Based QA uses a context document (say, a paragraph).
   b. Open Domain QA uses a huge corpus, finds a relevant context document itself, then does Context Based QA.

   Check SQuAD dataset.

2. **Text Retrieval (or Information Retrieval):**
   Is used for finding “relevant things”, which are similar with your “query”. You can apply this to many, many, areas.
   This task was industrially used before it was being approached with Machine Learning.
   Used in Google search engine, Shazam, genetics research, by copyrights firms…
   Also used in Open Domain QA to find context paragraphs.

Problems in NLP Notebook

3. Text Classification (we particularly do sentiment analysis in the notebook):

Text classification is incredibly popular in industry, and very easy to implement.
Can be at the beginning of many NLP pipelines. You sometimes may want to classify your text, before doing any other ML.

A perfect example is chatbots. Chatbots include several NLP modules in them (language generation, dialoge management etc.), however, most fundamental one of these modules is the intent classification model.

Kaggle might be the best place to get started.

4. Token Tagging (a popular application: Named Entity Recognition):

Again, very popular in the industry.
Might be used for morphological analysis of the sentence (noun verb etc.), finding named entities (Obama, Trump etc.)

Similar to text classification, might be at the beginning of your NLP product pipeline. You may want to tag the text first.

Problems in NLP Notebook

5. Translation

Used immensely in the industry, however, seems like a bit of a solved problem. (Google Translate etc.)

Qualitative challenges still exist though.

6. Speech:

A very large field, filled with a lot of signal processing theory.

Used in virtual assistants, and other kinds of digital agents.

Problems in NLP Notebook

7. Language Modelling (and language understanding)

When you use state of the art approaches, language modelling is at the core of all NLP related problems.

Understanding (modelling) language is a critical requirement to process language, in most cases.

Check BERT.

8. Text Generation:

Basically, every problem in natural language processing field can be modelled as a Text Generation task.
(text generation modelling will not always be the best solution for your problem though!)

It is the state of the art approach for question answering and translation problems.

Check GPT-3.

What are some things to deal with in NLP?

Let’s say we will do sentiment analysis on a sentence. This, is a binary classification for each sentence.

1. **How can I process a sequence in a model?**
2. **What type should the items (tokens) be in the sequence? Words? Letters? Subword?**
3. **Say we decided on tokenizing with the roots and with the additions (subwords) in a word. How do I learn what is a root and what is an addition?**
4. **How should I represent each token (subword)? One hot encodings? A way that gives me semantic info on the subword?**
5. **To obtain semantic info on subwords, what kind of a model can I use?**
6. **To train that model, what kind of a task can I formulate?**

What are some things to deal with in NLP?

Let's say we will do sentiment analysis on a sentence. This, is a binary classification for each sentence.

1. **How can I process a sequence in a model?**
2. **What type should the items (tokens) be in the sequence? Words? Letters?**
3. **Say we decided on tokenizing with the roots and with the additions (subwords) in a word. How do I learn what is a root and what is an addition?**
4. **How should I represent each token (subword)? One hot encodings? A way that gives me semantic info on the subword?**
5. **To obtain semantic info on subwords, what kind of a model can I use?**
6. **To train that model, what kind of a task can I formulate?**

Answers:

1. **RNN, LSTM, or Transformer models.**
2. **Subword Tokenization is currently the most popular way.**
3. **SentencePiece (2018) algorithm by Google is a popular way.**
4. **First, represent with one hot encodings, then use a lookup matrix for the embedding of the word.**
5. **BERT based models are currently the most popular. Word2Vec is the easiest to understand.**
6. **Skipgram is what Word2Vec uses. BERT based models use Masked Language Modelling.**

Concepts on NLP

1. **Preprocessing**
   a. Corpus, Tokenizers, Special Tokens, Lowercasing, Stopword and Punctuation Cleaning, Stemming and Lemmatization
2. **Obtaining Word Features**
   a. Vector Space Models (Embedding Layers), N-grams, Word Counts, Vocabulary
3. **Models and Algorithms**
   a. Transformers (and attention), Sequential Models (RNN, GRU, LSTM), Cosine Similarity, Edit Distance
   b. Logistic Regression, Bayesian Models, kNN algorithm, Hash Tables, Index Search Algorithms, Hidden Markov Models, Viterbi Algorithm, Dynamic Programming
4. **Problems**
   a. Text Retrieval (Search), Named Entity Recognition, Machine Translation, Part of Sentence Tagging,
5. **Visualization**
   a. PCA, tSNE
6. **Evaluation**
   a. Perplexity, BLEU, ROGUE
7. **Experimentation and Development**