NLP Project Plan

Sentiment Analysis Feature Extraction

* Stemming / Lemmitisation
* Part of Speech tagging
* Word-sense disambiguation – SentiNet
* Word chunks
* N-grams
* Negation

Week 6 - Unsupervised Learning

* Word vector clustering
* Feature weighting / dimensionality reduction important
* K-means –
  + linear in complexity in all relevant factors: number of iterations, number of clusters, number of document vectors and dimensionality of the space.
  + The total vocabulary used to create the feature space, on the other hand, even after stopwords removal will run up to tens of thousands of unique words - realistically, data will have a very small number of dimensions filled with actual occurrence numbers, while the rest will contain zeros
* Single value decompositiion algorithm (SVD) – aims to significantly reduce data dimensionality to help expensive algorithms like clustering deal with it more efficiently, while keeping as much of the valuable information in the reduced data as possible – apply before k means
* Evaluation –
  + Purity: sum(majority cluster count per cluster) / N
  + Homogenetity (precision)
  + Completedness (recall)
  + V-Measure (F1 Score)
* **Topic modelling** - topic modelling assumes the existence of an underlying generating process behind the document composition, that is responsible for the selection of topics and generation of words representing these topics in the documents observed in the collection. The goal of the algorithm is to reverse-engineer the generation process and to estimate the parameters that drive this process, based on the observations from the data. - Since topic models can be applied to any amount of unlabelled data, they are very useful in organising data, especially in large collections, where manual inspection and annotation are impossible, and in providing insights about the underlying thematic structure of large collections of unstructured data - Topic modelling is a text-mining technique, which aims to discover abstract “topics” that occur in a collection of documents. It relies on the idea that documents on a particular topic would contain topic-related words in a particular, representative proportion: for instance, “car” and “driving” will occur more often in documents about cars; “phone” and “keyboard” in documents about electronics;
  + Latent Dirichlet Allocation - relies on the intuition that in any collection the documents normally cover a small number of topics and that topics often use a small number of words. Thus, LDA introduces sparse Dirichlet prior distributions over document–topic and topic–word distributions.
  + As with the other unsupervised algorithms, the data is not labelled, i.e., the algorithm has no prior knowledge of the existing “topics” or words representing such topics. The “topics” are abstract concepts behind the words used in documents, yet because of the probabilistic nature of the word use, the inferred hidden structure uncovered by the algorithm normally resembles the thematic structure of the collection.
  + LDA provides you with a much richer hidden topic structure, therefore comprehensive evaluation of topic modelling algorithms is a question actively researched in NLP and ML

Week 7 - Sequence Modelling and structured prediction

* A Markov chain is a model that defines for us the probabilities of sequences of random variables – states, which can take on values from some predefined set
* PoS tagging can be modelled using an extension to Markov Models – Hidden Markov Models (HMM), that allow you to take into account both observed (words) and hidden (PoS tags) events
* Languange model
* A Hidden Markov Model (HMM) helps you capture both observed events (e.g., words that you see in the input) and hidden events (e.g., part-of-speech tags) -  One particular problem that HMM faces is its inability to take into account previously unseen events (e.g., unknown words)
* HMM Tagging/Decoding – using Bayes to predict with HMM assumptions
* Viterbi Algorithm for Decoding - find the most probable tag sequence that generated the observation sequence “John carried a tin can” and estimate its probability. You can calculate the probabilities manually or, alternatively, you can implement a Viterbi algorithm in Python to solve this task. -  HMM parameterised with transition (defining the transitioning from the previous to the current states) and emission (defining the emission of observations from states) probabilities
* Named entitiy recognition (NER) –
  + PER (people) – for people, characters, and similar.
  + ORG (organisation) – for companies, organisations, sports teams, etc.
  + LOC (location) – for regions, mountains, rivers, seas, etc.
  + GPE (geopolitical entities) – for countries, states, and similar.
  + IO, BIO, and BIOES taggings
* Alternative to HMM - An alternative to relying on such a generative model would be building a model that can help the PoS tagger assign the correct PoS tag by incorporating various features pertaining to unknown words themselves (e.g., words ending with -ed have a high chance of being verbs) or to the surrounding words (e.g., words following the or other determiners have a high chance of being nouns). Discriminative sequence models based on log-linear models allow one to incorporate such features. In this lesson, you will look into one such model, the linear chain conditional random field (CRF)
  + At each time step, the CRF computes log-linear functions over the set of relevant features, and then aggregates these local features to produce the global probability for the whole sequence. Just like classifiers in the supervised machine learning setting that you looked into in Week 4, the CRF builds a function f that maps the features of the input token x to the output label y; the difference is that in the case of the CRF, it maps global features of the entire input sequence X to the entire output sequence Y.]

**Semantics**

* Lexical semantics -  The two instances bank1and bank2 , which have the same lemma but quite distinct senses, are called **homonyms**, and the relation between two senses is called **homonymy**.
  + I went to the bank1 to withdraw money.
  + We sat on the bank2 and had a picnic
  + bank3 in blood bank3is an extension of the sense of bank1 to a different domain. Both these senses denote repositories of entities: while bank1 refers to the financial domain, bank3 refers to the biological domain. In such cases, the relation between senses is called **polysemy**
  + This subtype of polysemy relation is called **metonymy**, and it applies to the cases where one aspect of a concept or an entity is used to refer to the other aspects of the entity or to the entity itself. Table 1 shows some other common examples of metonymy:
    - Author (Shakespeare wrote Hamlet)
    - Works of author (I studied Shakespeare at school)
* Relation between word senses –
  + **Synonymy** holds between the senses of two words when these senses are identical or nearly identical. For instance, words like couch and sofa -  Another aspect to keep in mind is that synonymy holds between word senses rather than words: for instance, while big and large are synonymous when used in some senses (e.g., meaning(big1 room) = meaning(large room)), in other senses they are not synonymous (e.g., meaning(big2sister) ≠ meaning(large sister))
  + **Antonymy** holds for words with the opposite meaning (for example, big and little
  + (car, dog, cherry) is called a **hyponym**, and the superordinate term (vehicle, animal, fruit) is called a **hypernym**. Hyponymy is an important relation in language as it helps you organise the world into clear classification, or a taxonomy. -  Socrates is a man. All men are mortal. Therefore, Socrates is mortal. Here, the superordinate term, man, has a certain characteristic (of being mortal) that applies to all instances of the type and, therefore, applies to a specific instance as well. This is called entailment, and it is a challenging task within NLP.
  + Semantic fields allow you to put together various entities and properties that relate to entire sets of words from a single domain: for instance, {*reservation*, *flight*, *booking*, *meal*, *plane*, *price)*
* WordNet is the largest, most comprehensive and most widely used database of lexical relations. WordNet is structured hierarchically
* Vector based semantics -  representation of word meaning from the computational, data-driven perspective. - , is the ability of an algorithm to not only detect when certain words are synonymous in a particular sense or context (e.g., “excellent” and “brilliant”) but also to what extent they are semantically similar. - distributional semanticst to be used
  + Distributional semantics uses the same principle: similar words have similar vectors because they tend to occur in similar documents or similar contexts. This means that to build a word vector you can “flip” the dimensions of the document–term matrix and represent the meaning of a word by the documents it tends to occur in. Alternatively, you can build a term–term matrix.
  + select a predefined number of dimensions (commonly between 10,000 and 50,000) to represent the most frequent context words excluding stopwords.
  + Based on the co-occurrence counts from the table in figure 5, you can see that words cherry and strawberry are more similar to each other: leaving the differences in the absolute number of occurrences aside, both words have higher co-occurrence counts with the context words like pie and sugar than with computer and data. - note that in practice no target word would co-occur with a number of words even remotely close to 10,000, meaning that even after context word selection, the co-occurrence matrix would still be very sparse. Dimensionality reduction techniques, such as singular value decomposition, which you used for clustering in Week 5, are commonly applied to help alleviate this issue.
  + tf-idf weighting,  can help with the above or positive pointwise mutual information (PPMI), which draws on the intuition that the best way to weigh the association between two words is to ask how much more the two words co-occur in the data than you would have expected them to occur by chance, i.e., if they were independent - The PMI values range from negative to positive infinity, however, the negative values simply show to what extent the words are unrelated, which is not the main point of the measure focused on estimating how much the words are actually related. Therefore, in practice, it is more common to use Positive PMI (PPMI), with replaces all negative values with zero:
  + distributional word representations are the earliest examples of representation learning  - finding self-supervised ways of automatically learning useful representations of the input text instead of creating representations by hand via feature engineering
* Vector based semantic composition
  + build a representation of phrases, e.g., “excellent movie” or “brilliant film” and measure their similarity? -  distributional approach is not scalable for longer phrases:
  + explore the compositionality of individual units, i.e., words, in longer phrases, and build vectors for phrases and sentences using -n distributional word vectors as building blocks. - compositional semantics.
* Word embeddings – smaller dense word vectors - short dense vectors: for instance, representing words with a few hundred of dimensions instead of thousands requires classifiers to learn far fewer weights, and the smaller parameter space helps improve generalisation and avoid overfitting -  a disadvantage of word embeddings as compared to distributional vectors is that the dimensions no longer have a clear interpretation as they do not explicitly correspond to a particular word from the vocabulary.
  + Word2Vec
    - skip-gram with negative sampling - Note that the skip-gram model makes a simplifying assumption of context words independence here (you have seen such assumptions applied before in the context of Naïve Bayes and Markov models)
      * word2vec performs binary classification training a logistic regression classifier. Here is the procedure that the skip-gram model follows:
        + It treats the target word *w* and a neighbouring context word *c* as positive examples.
        + Then, it randomly samples other words in the vocabulary and uses them as negative samples.
        + Next, it uses logistic regression to train a classifier aimed at distinguishing between positive and negative samples.
        + Finally, it uses the learned weights as the word embeddings.
      * the model aims to maximise the dot product (i.e., similarity) of the word w with the actual context words cpos, and, at the same time, minimise the dot products of the word w with the k negative samples of non-neighbour words cneg. In practice, this loss function is minimised using stochastic gradient descent (which is commonly used to optimise the loss functions in a number of models, including logistic regression and neural networks)
      * The skip-gram model uses negative sampling, which means that it randomly selects a number of negative examples with the ratio of negative-to-positive examples defined by the parameter k: in other words
      * step of gradient descent. The skip-gram model tries to shift embeddings so that the target embeddings (i.e., for apricot here) move closer to (and result in a higher dot product with) context embeddings of the actual nearby words (e.g., jam) and further away
    - continuous bag-of-words model
    - GloVe – to research - static embedding models like word2vec - methods learn one fixed embedding for each word in the vocabulary regardless of different contexts of use and subtleties in meaning
* Semantic evaluation –
  + extrinsic evaluation can measure how the semantic representations component affects the results of the overall task. In contrast, intrinsic evaluation is concerned with assessing the quality of the semantic representations themselves.
  + SimLex-999 / TOEFL dataset
  + Stanford Contextual Word Similarity (SCWS) and the Word-in-Context (WiC)
  + compositional abilities of semantic representations (i.e., their abilities to model meaning beyond individual words)
  + Analogy using word vectors comparisons
  + Evaluation to find semantic similarity, with and without context, to those evaluating algorithms’ ability to capture relational meanings and to solve analogy tasks

**Neural Language Models - feedforward**

* The module on Neural Language Models delves into the intricacies of n-gram language models and their neural counterparts, highlighting the limitations of the former due to reliance on specific contexts seen during training. For instance, traditional models might fail to predict the word "fed" following "dog" if only "cat gets fed" was encountered during training, despite the semantic similarity between "dog" and "cat." Neural models, leveraging word embeddings, excel in generalizing based on semantic similarities, thereby significantly enhancing predictive accuracy over unseen data.
* Further exploration into building and training neural language models reveals that they utilize feedforward networks to process input from preceding words and output probability distributions for possible next words. These networks consist of input, hidden, and output layers, with hidden units performing computations on inputs using weighted sums and non-linear activation functions. The architecture facilitates links between all units across adjacent layers, ensuring a fully connected network. Crucially, the feedforward model approximates the probability of a word given the context by utilizing word embeddings for representation, enabling predictions based on semantic relationships rather than exact word tokens. This approach allows for more accurate and generalizable language modeling, overcoming the limitations of traditional n-gram models.

To summarise, here is the procedure followed by the feedforward LM:

1. First, you take the embeddings for the *n-context* words and concatenate them.
2. Then, you multiply this vector by *W* and add *b*, and you then pass the output through the non-linear activation function to get the hidden layer *h*.
3. Next, you multiply *h* by *U*.
4. Finally, you apply softmax, after which each node*i* in the output layer estimates the probability *P*(*wt= i*| *wt-1*, *wt-2*, *wt-3*).

* Training for a neural language model typically proceeds by taking as input a very long text, concatenating all the sentences, starting with random weights, and then iteratively moving through the text predicting each word *wt*. The weights are optimised using gradient descent and backpropagation, and the goal of the procedure is to make sure that the model prediction for the output words is as close to the actually observed words in the data as possible.
* Thus, neural language models use a neural network as a probabilistic classifier, to compute the probability of the next word given the previous *n*words. You can use pretrained embeddings (such as word2vec, outlined in Week 7) as a component of the neural LM, or you can initialise input embeddings randomly and learn them on the specific task in the process of training the LM on the data.

**Neural Language Models - recurrent neural network** (**RNN**).

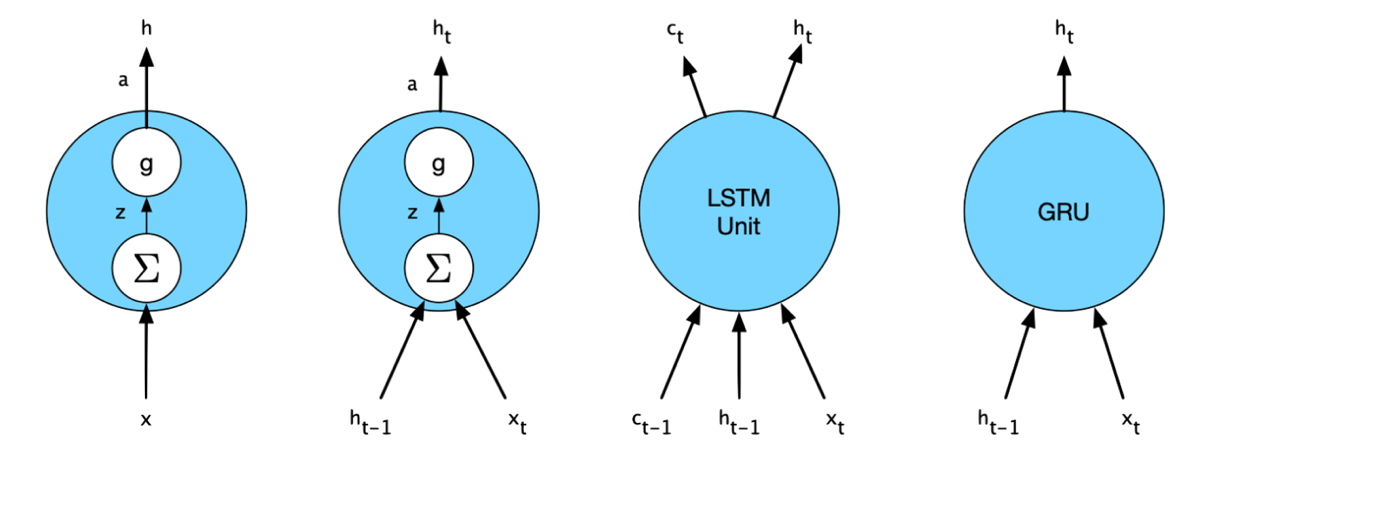
* This term is used for any network that contains a cycle within its connections, i.e., where the value of a unit is directly, or indirectly, dependent on its own earlier outputs as an input.
* An RNN shares many properties with a simple feedforward network overviewed before: just like before, the current input xt is first multiplied by a weight matrix W, then passed through a non-linear activation function, e.g., σ, to compute the values for a layer of hidden units, which is then used to calculate the output yt. The key difference with the earlier, simple feedforward models is, however, the fact that input to the hidden layer is expanded with the value of the hidden layer from the preceding point in time, ht-1 – this is illustrated in Figure 3 with a dashed line. Essentially, this means that by taking as input the value of the hidden layer from the previous timestamp, the network keeps some sort of memory or context, which affects the current decision, and this also allows the network to learn from information outside the precise and fixed-length context window.
* This combination of an RNN and a feedforward classifier is a good example of a deep neural network – a network consisting of multiple hidden layers. In addition, during training such a network uses the loss function from a downstream application (e.g., sentiment analysis) to adjust the weights all the way through the network – such training regimes are called end-to-end training.

**Neural Language Models - RNN - LSTM**

* While it is straightforward to assign a high probability to the word was after the local context “the student”, assigning an appropriate probability to were is difficult: first of all, the relevant context (“the books”) is quite distant, secondly, the more immediate context (“the student”) is singular. Ideally, the network would need to store the relevant information about “the books”, while processing the intermediate parts.
* The second problem with training RNNs arises from the need to backpropagate the error signal back through time and multiple layers. In practice, this involves many multiplication steps, which result in gradients being driven to zero – this problem is widely acknowledged in deep learning research as the vanishing gradients problem. To address the problems with keeping the most relevant past context, two popular architectures – long short-term memory (LSTM) and gated recurrent units (GRU) were introduced, which help the network to better handle the information from the local as well as distant context.
* Specifically, LSTMs divide the context management problem into two sub-problems: removing (forgetting) information from the context that is no longer needed and adding information that is likely to be needed for later decision-making. For that, LSTM adds an explicit context layer to the architecture – thus, it contains both the usual recurrent hidden layer and the new context layer. In addition, it employs specialised neural units that make use of gates to control the flow of information into and out of the units. Each gate consists of a feedforward layer, followed by a sigmoid activation function, followed by a pointwise multiplication with the layer being gated. The choice of the sigmoid function here is well-motivated: since it pushes its outputs to either 1 or 0, this works as a definitive choice on whether to keep (add) the information in the layer or not.

**Neural Language Models - RNN – GRU**

* LSTM hard to train – Gated Recurrent Units (GSU) use 2 gates –



**Transformer models – to research**

**Structure - GPT**

**1. Task and Data Analyses**

- Introduction - Briefly introduce the assignment, its objectives, and its significance.

- Task Analysis:

- Describe the opinion mining task.

- Break down the task into constituent parts.

- Outline the system-level approach for your solution.

- Data Analysis:

- Explore and describe the dataset provided.

- Discuss any notable characteristics, distributions, or potential challenges the data presents.

- Theoretical Background:

- Review relevant literature and theories that underpin your approach.

- Cite references that informed your system design and methodology.

**2. Data Preprocessing**

- Data Ingestion:

- Detail the process of importing and handling the raw data.

-Data Cleaning and Transformation:

- Describe any cleaning steps to remove noise or irrelevant information.

- Explain transformation processes to convert data into a usable format.

- Preprocessing Techniques:

- Justify the selection of NLP preprocessing techniques applied (e.g., tokenization, normalization).

- Discuss the rationale behind each preprocessing decision.

**3. Product Feature Extraction**

- Methodology:

- Describe the methods used for identifying key phrases that signify product features.

- Explain the role of PoS tagging, chunking, or parsing in your approach.

- Algorithm Implementation:

- Detail the algorithms implemented for feature extraction.

- Provide theoretical justifications and references for the chosen algorithms.

- Feature Candidate Pruning/Ranking:

- Discuss any techniques used to refine the set of product features identified.

**4. Sentiment Analysis**

- Sentiment Detection:

- Outline how sentiments are identified within the data.

- Algorithm Selection and Justification:

- Describe the algorithm(s) used for sentiment analysis.

- Justify your choice based on theoretical and practical considerations.

- Comparative Experiment:

- Compare two significantly different algorithms or feature sets.

- Formulate a theoretical hypothesis for the comparison.

- Present a thorough analysis and discussion of the comparative results.

**5. Evaluation and Discussion**

- Evaluation Methodology:

- Define the metrics and procedures used to evaluate each subsystem and the entire pipeline.

- Explain the rationale behind your evaluation strategy.

- Results Reporting:

- Present the results of your implementations, including precision, recall, and any other relevant metrics.

- Include representative examples of opinion "summaries".

- Discussion:

- Analyze performance trends observed.

- Discuss any challenges encountered and how they were addressed.

-Conclusions and Future Work:

- Summarize key findings and their implications.

- Suggest areas for future research or improvements to the opinion miner.

**Task and Data Analysis**

**Data Pre-processing**

* Segmentation / Tokenisation
* Frequency Analysis
* Morphology - Stemming / Lemitisation

**Product Feature Extraction**

* Term weighted processing
  + Document frequency / inverse document frequency
  + TF-IDF score
* Remove stop words and punctuation
* Part of Speech tagging
* Word-sense disambiguation – SentiNet
* Word chunks
* N-grams
* Negation
* Parsing
  + Early parsing algorithm
  + Top-down / bottom-up parsing
* Grammatical relations
* Non-negative matrix factorisation (NMF)
* Neural network approaches

**Sentiment Analysis**

* Unsupervised models
  + Word vector clustering
  + Feature weighting / dimensionality reduction important
  + K-means
  + Single value decompositiion algorithm (SVD)
  + Topic modelling -*a text-mining technique, which aims to discover abstract “topics” that occur in a collection of documents*
    - Latent Dirichlet Allocation
* A Markov chain is a model that defines for us the probabilities of sequences of random variables
* PoS tagging can be modelled using an extension to Markov Models – Hidden Markov Models (HMM)
* Languange models
  + A Hidden Markov Model (HMM) helps you capture both observed events (e.g., words that you see in the input) and hidden events (e.g., part-of-speech tags)
  + HMM Tagging/Decoding – using Bayes to predict with HMM assumptions
  + Viterbi Algorithm for Decoding -
  + Named entitiy recognition (NER)
* Discriminative sequence models
  + linear chain conditional random field (CRF)
* Lexical semantics -
  + WordNet
  + Vector based semantics - distributional semantics
    - tf-idf weighting
    - positive pointwise mutual information (PPMI)
  + Representation learning
  + Vector based semantic composition - compositional semantics.
  + Word embeddings
    - Word2Vec
    - skip-gram with negative sampling
    - Continuous bag-of-words model
    - GloVe

**Evaluation and Discussion**

* Mean reciprocal rank
* Semantic evaluation
  + extrinsic / intrinsic
  + SimLex-999 / TOEFL dataset
  + Stanford Contextual Word Similarity (SCWS) and the Word-in-Context (WiC)
  + compositional abilities of semantic representations (i.e., their abilities to model meaning beyond individual words)
  + Analogy using word vectors comparisons
  + Evaluation to find semantic similarity, with and without context, to those evaluating algorithms’ ability to capture relational meanings and to solve analogy tasks
* Unsupervised evaluation
  + Purity: sum(majority cluster count per cluster) / N
  + Homogenetity (precision)
  + Completedness (recall)
  + V-Measure (F1 Score)
* Supervised evaluation
  + Accuracy
  + Precision
  + Recall
  + F1 score