Section 1 : Semantics

- Question 1A

Consider the following sentences.

(1) - The black horse jumped over the white fence
(2) - The black fence was too high for the black horse
(3) - The white horse could not have jumped over the black fence
(4) - The black horse jumped successfully over the black fence
(5) - The white fence was also too high for the white horse

Which is the similarity score of 'white' and 'black'?

The contex window of two tokens for each sentence, gives us:

(1) black ~ the, horse, jumped
white ~ over, the, fence

(2) black ~ the, fence, was, for, the, horse

(3) black ~ over, the, fence
white ~ the, horse, could

(4) black ~ the, horse, jumped, over, the, fence

(5) white ~ the, fene, was, for, the, horse

Therefore, this gives us the following vectors:

	the	horse	jumped	over	fence	was	for	could
black	6	3	2	2	2	1	1	0
white	4	2	0	1	2	1	1	1

Let
$$\left.\begin{aligned} v_B &:= (6,3,2,2,2,1,1,0) \\ v_W &:= (4,2,0,1,2,1,1,1) \end{aligned}\right\}$$
the cosine similarity

can be calculated by:

$$\cos(v_B, v_W) = \frac{|v_B \cdot v_W|}{\|v_B\|\,\|v_W\|} \triangleq \frac{38}{\sqrt{59}\sqrt{28}} = \frac{38}{\sqrt{1652}} \approx 0.86$$

$$|v_B \cdot v_W| = 24 + 6 + 2 + 4 + 1 + 1 = 38$$
$$\|v_B\| = \sqrt{36 + 9 + 4 + 4 + 4 + 1 + 1} = \sqrt{59}$$
$$\|v_W\| = \sqrt{16 + 4 + 1 + 4 + 1 + 1 + 1} = \sqrt{28}$$

Therefore, we can say that 'white' and 'black' have a similarity score of 0.86.

• Question 1B

How can dependency parsing be used in a vector space model?

Instead of restricting the vectors to the count of occurrences of each word in the context window, we can extend this by analysing the syntactic dependencies of each word in the context. In this way, this extra information will let us know if for instance words from the context are interacting with the verb or subject. If we are able to differenciate more the words from the context we should be able to have more information on the target words.

• Question 1C

What is the opposite of a hyponymy?

A hyponomy is a word whose semantic field is included in another word. The opposite of a hyponymy is a hypernym. While hyponomy refers to the specific word hypernym refers to the generic one.

- Question 1D

Give to examples of semantic roles and construct a sentence in which both of these can be used. Annotate the relevant words in the sentence for each of these two roles.

$\underbrace{\text{The pilot}}_{\text{Agent}}$ flew $\underbrace{\text{from Barcelona}}_{\text{Source}}$

Agent: deliberately performs the action

Source: wher the action originated

Section 2 : Part-of-speech tagging

- Question 2A

By using Viterbi Algorithm, find the most likely sequence of tags for 'tag this text'.

$w_0 = START$, $w_1 =$ tag, $w_2 =$ this, $w_3 =$ text

Using viterbi algorithm, the first iteration can be directly calculated:

- Iteration = 0

$$\pi_{tags,0} = \begin{array}{c} \begin{array}{cccc} START & N & V & O \end{array} \\ \left[\begin{array}{cccc} 1.0, & 0, & 0, & 0 \end{array}\right] \\ \begin{array}{cccc} [S] & [N] & [V] & [O] \end{array} \end{array} = [\pi_{START,0}, \pi_{N,0}, \pi_{V,0}, \pi_{O,0}]$$

- Iteration = 1, $\pi_{tags,1}$

As $P(S|\cdot) = 0$, we have that $\pi_{START,i} = 0 \quad i = 1, 2, 3.$

$$\pi_{N,1} = P(W_1|N) \cdot \max \begin{cases} \pi_{S,0}\, P(N|S) \\ \cancel{\pi_{N,0}}^{\,0}\, P(N|N) \\ \cancel{\pi_{V,0}}^{\,0}\, P(N|V) \\ \cancel{\pi_{O,0}}^{\,0}\, P(N|O) \end{cases} = 0.2 \cdot 0.4 = 0.08$$

As $\pi_{N,0} = \pi_{V,0} = \pi_{O,0} = 0$:

$$\pi_{V,1} = P(w_1|V) \cdot \pi_{S,0}\, P(V|S) = 0.6 \cdot 0.5 = 0.3$$

$$\pi_{O,1} = P(w_1|O)\, \pi_{S,0} \cdot P(O|S) = 0.1 \cdot 0.1 = 0.01$$

Therfore $\pi_{tags,1} = [0, 0.08, 0.3, 0.01]$

$$\underset{[S,N]}{0.08}\quad \underset{[S,V]}{0.3}\quad \underset{[S,O]}{0.01}$$

Let's calculate $\pi_{tags,2}$.

$$\pi_{N,2} = \underset{0.2}{P(w_2|N)} \cdot \max \begin{cases} \pi_{N,1}\, P(N|N) = 0.08 \cdot 0.4 = 0.032 \\ \pi_{V,1}\, P(N|V) = 0.3 \cdot 0.3 = 0.09 \\ \pi_{O,1}\, P(N|O) = 0.01 \cdot 0.4 = 0.04 \end{cases}$$

$\Rightarrow \pi_{N,2} = 0.09 \cdot 0.2 = 0.018$

$$\pi_{V,2} = \underset{0.1}{P(w_2|V)} \max \begin{cases} \pi_{N,1}\, P(V|N) = 0.08 \cdot 0.2 = 0.016 \\ \pi_{V,1}\, P(V|V) = 0.3 \cdot 0.2 = 0.06 \\ \pi_{O,1}\, P(V|O) = 0.01 \cdot 0.3 = 0.03 \end{cases}$$

$\Rightarrow \pi_{V,2} = 0.06 \cdot 0.1 = 0.006$

$$\pi_{O,2} = \underset{0.7}{P(w_2|V)} \max \begin{cases} \pi_{N,1}\, P(O|N) = 0.08 \cdot 0.4 = 0.032 \\ \pi_{V,1}\, P(O|V) = 0.3 \cdot 0.5 = 0.15 \\ \pi_{O,1}\, P(O|O) = 0.01 \cdot 0.3 = 0.003 \end{cases}$$

$\pi_{O,2} = 0.15 \cdot 0.7 = 0.105$

This is $\pi_{Tags,2} = [0, 0.018, 0.006, 0.105]$
$[S,V,N]\ [S,V,V]\ [S,V,O]$

Finally:

$$\pi_{N,3} = \underset{0.5}{P(w_3|N)} \cdot \max \begin{cases} \pi_{N,2}\, P(N|N) = 0.018 \cdot 0.4 = 0.0072 \\ \pi_{V,2}\, P(N|V) = 0.06 \cdot 0.3 = 0.018 \\ \pi_{O,2}\, P(N|O) = 0.105 \cdot 0.4 = 0.042 \end{cases}$$

$$\pi_{V,3} = \underset{0.1}{P(w_3|V)} \max \begin{cases} \pi_{N,2}\, P(V|N) = 0.018 \cdot 0.2 = 0.0036 \\ \pi_{V,2}\, P(V|V) = 0.006 \cdot 0.2 = 0.0012 \\ \pi_{O,2}\, P(V|O) = 0.105 \cdot 0.3 = 0.0315 \end{cases}$$

$$\pi_{O,2} = \underset{0.1}{P(w_3|O)} \max \begin{cases} \pi_{N,2}\, P(O|N) = 0.018 \cdot 0.4 = 0.0072 \\ \pi_{V,2}\, P(O|V) = 0.06 \cdot 0.5 = 0.03 \\ \pi_{O,2}\, P(O|O) = 0.105 \cdot 0.3 = 0.0315 \end{cases}$$

This is
$$\pi_{N,3} = 0.5 \cdot 0.042 = 0.021$$
$$\pi_{V,3} = 0.1 \cdot 0.0315 = 0.00315$$
$$\pi_{O,3} = 0.1 \cdot 0.0315 = 0.00315$$

$\pi_{Tags,3} = [0, 0.021, 0.00315, 0.00315]$
$[S,V,O,N]\ [S,V,O,V]\ [S,V,O,O]$

This means that the most likely tag is
[S, V, O, N] with a probability of 0.021.

- Question 2B

What is an advantadge of the Viterbi Algorithm over an exhaustive search for all possible PoS tag combinations?

Viterbi Algorithm uses dynamic programming. An algorithm that has to check each possible combination of PoS tags would take an exponential time. However, be keeping results in memory, Viterbi is able to accomplish that in $O(Tn^2)$ where T is a sequence of observations from a HMM with n states.

- Question 2C

Give one advantadge and disadvantadge of unsupervised learing of a HMM by means of Baum-Welch as opposed to supervised learning of the model. Why might you wish to combine supervised and unsupervised learning for HMM?

- It's not easy to find datasets with annotations to use for supervised learning. Baum-Welch provides a way of calculating the emission and transition probabilities of an HMM given just the number of states.

- We might wish to combine supervised and unsupervised learning for HMM if we want as unsupervised does not give much information about the tags that are found giving more understanding and information about the results.

Section 3: Sentiment Analysis

Consider the following sentences from product reviews with sentiment scores, using a range of -1, 0, 1.

-1 The food was great, but the service in this restaurant was unfriendly

+1 The steak was cooked to perfection and the service was great

0 The best part of the meal was beer

- Question 3A

Create a sentiment lexicon on the basis of the review sentences given above

- Words that we see related to the negative review:

  great, unfriendly

- Words that we see related to the positive review :

  perfection, great

- Words that we see related to the neutral review:

  best

Therefore, our sentiment lexicon is going to include: great, unfriendly, perfection and best.

While there are many ways to find scores for a categorical or numerical sentiment lexicon I have decided to use the following approach.

- Assign to each word the average score from the sentences that it belongs. In this way we have:

great → $\frac{-1+1}{2} = 0$

unfriendly → -1

perfection → 1

best → 0

⇒

Word	Score
Perfection	1
Great	0
Best	0
unfriendly	-1

• Question 3B

Recall that a count vector represents the proportions of negative and positive sentiment in a review text. Using the sentiment lexicon you created in question 3A, calculate the count vector for each review sentence given above.

As we have 3 values in this sentiment lexicon my count vector will have 3 components positive, neutral, negative.

$$r_1 = [0, 0.5, 0.5]$$
$$r_2 = [0.5, 0.5, 0]$$
$$r_3 = [0, 1, 0]$$

• Question 3C

Give an example of aspect-based sentiment analysis from the review sentences given above

When thinking about ABSA we need to take into account the different categories of the product or service that we are analysing. A clear example can be found in the first review:

- 1 : The food was great, but the service on the restaurant was unfriendly.

Even though the overall score is -1, the customer is happy with the food. This can be seen using ABSA:

	Score
Food	0
Service	-1

Section 4: Information Extraction & Knowledge Graph

Consider the following sentence:

'David Robert Joseph Beckham is a former professional footballer who played for Manchester United and the England national team'

- Question 4A

Annotate the sentence above for PER, ORG and LOC by the use of the IOB tagging scheme. Explain the reasoning behind your annotations:

David	Robert	Joseph	Beckham	is	a	former	professional	footballer
B-PER	I-PER	I-PER	I-PER	O	O	O	O	O

who	played	for	Manchester	United	and	the	England
O	O	O	B-ORG	I-ORG	O	O	B-ORG

national	team.
I-ORG	I-ORG

Any word not related to PER / ORG / LOC can be categorised as 'Outside' (O).

England national team can be thought as well as England national team but because of the context I feel that the given answer is more approprate.

I-LOC O O

- Question 4B

The sentence above provides a positive instance for extracting the 'played for' relation between a 'person' (footballer) and an 'organization' (football team). Give a negative instance of this relation.

That chess master plays for money.

There is no relation with the chess master and money in this example.

- Question 4C

How can clustering be used in taxonomy extraction? What are potential problems with this approach?

By using Unsupervised Term Extraction, we can have taxonomy extraction. The main idea is to find different clusters with the NP of the texts provided. ~~However~~, NP may have different ratings, measures such as PMI can help to identify the ratings between them.