COMP 6721 Applied Artificial Intelligence (Winter 2022)

Lab Exercise #10: Introduction to NLP

Question 1 Assume that we are working with the Shloutan language. If you don't know Shloutan, don't worry; it is a simple language made of only 5 words: loola nikee aloka bibi vo.

You want to build a word language model for Shloutan. The training corpus that you use is the following:

"Loola nikee. Aloka bibi vo. Vo bibi loola. Loola nikee bibi vo. Vo. Vo. Aloka bibi loola. Loola aloka aloka. Loola loola. Nikee nikee nikee. Bibi vo. Bibi vo. Vo Vo. Nikee loola."

You can ignore case distinctions and sentence boundaries when answering the following questions.

(a) What is the value of P(vo | bibi)?

(b) What is the value of P(bibi vo)?

(c) Build a bigram language model based on this training corpus. Show the frequencies and the probabilities for each bigram.

Frequencies

	loola	nikee	aloka	bibi	vo
loola					
nikee					
aloka					
bibi					
vo					

Conditional Probabilities

	loola	nikee	aloka	bibi	vo
loola					
nikee					
aloka					
bibi					
vo					

(d) Smooth your bigram language model using "add 0.5". Show the frequencies and the probabilities for each bigram.

Frequencies

	loola	nikee	aloka	bibi	vo
loola					
nikee					
aloka					
bibi					
vo					

Conditional Probabilities

	loola	nikee	aloka	bibi	vo
loola					
nikee					
aloka					
bibi					
vo					

(e) Using each language model from parts (c) and (d), which of the following two sentences is more probable. Show all your work.

sentence 1: Aloka vo nikee aloka.

sentence 2: Vo nikee nikee aloka.

Question 2 Consider the following context-free grammar:

S → NP VP	N → lion
VP → V	N → knife
VP → V NP	N → zoo
VP → VP PP	V → attacked
NP → DET N	PN → Jane
NP → PN	DET → the
NP → NP PP	DET → a
PP → PREP NP	PREP → with
	PREP → in

(a) Generate all possible parse trees for the following sentence:

    i *Jane attacked a lion with a knife*

Question 3 Consider the following sentences as training set for disambiguating the sense of the word *light*.

Sentence	Sense
There is a ray of light even if it is dim.	Sense1
This faint ray of light is barely enough to see.	Sense1
It needs to be light in weight to be portable.	Sense2
I can see dim red light coming from that room.	Sense1
This laptop is light enough to carry around.	Sense2
I am not hurt since I was hit by something as light as a candle.	Sense2
There is a faint hint of light at the end of the street.	Sense1
But the room was bright and light from the candle was warm as it fell on the surroundings.	Sense1
My bright red overcoat is light but fairly warm.	Sense2
This red candle stand is light enough to carry it to the room.	Sense2

Further consider this list of stop words:
> *a, and, are, as, at, be, for, from, in, is, it, if, of, that, this, the, to, was*

Using a Naive Bayes approach with:

- a context window of ± 3 words,
- a vocabulary with the size of 31,
- smoothing with the value of 0.2, and
- stop-word removal

Calculate the scores of each possible sense and find the most probable sense of the word *light* in the following sentence:

> *In her dark red room, that dim light from the candle was warm enough.*

Question 4 Preprocessing text is a very crucial step for any NLP related task. There are many frameworks (GATE[1], UIMA[2]) and libraries (CoreNLP[3], NLTK[4], spaCy[5]) available out there to facilitate text analysis.

Today we will be looking into spaCy to get you started with the basic steps.

Installing spaCy. Depending on your environment, use *one* of the following: pip: Before installing spaCy lets first make sure your pip and setuptools are up-to-date and install them:

```
pip install -U pip setuptools wheel
pip install -U spacy
```

conda:

```
conda install -c conda-forge spacy
```

After the installation, execute the following command to verify that you have installed spaCy version 3 or upper.

```
python -m spacy info
```

Downloading language models. Not all languages are preprocessed in the same way. Depending on the language, tokenization, sentence splitting, POS tagging, etc. varies. To facilitate analyzing different languages and genres, spaCy provides different pre-trained language pipelines to work with.[6] For starters, we will be working with the basic English model. We can download the model with the following command:

```
python -m spacy download en_core_web_sm
```

Let's start coding! Use the following to import the spaCy library and load the language model:

```
import spacy
nlp = spacy.load("en_core_web_sm")
```

Now let's try to preprocess the text in your worksheet and check if the POS tags we assigned are what spaCy returns as well:

```
doc = nlp("I prefer a direct flight to Chicago.")
```

---

[1] https://gate.ac.uk/

[2] http://uima.apache.org/

[3] https://stanfordnlp.github.io/CoreNLP/index.html

[4] https://www.nltk.org/book/

[5] https://spacy.io/usage/spacy-101

[6] https://spacy.io/usage/models

Depending on the language model and sentence provided as input, spaCy outputs an object doc, in our case with a variety of annotations encoded within. Linguistic annotations are available as token attributes. Let's try to print some basic attributes:

```
for token in doc:
    print(token.text, token.lemma_, token.tag_, token.dep_)
```

(a) Now try to process the following paragraph to see how spaCy performs sentence splitting:

> *"Superman was born on the planet Krypton and was given the name Kal-El at birth. As a baby, his parents sent him to Earth in a small spaceship moments before Krypton was destroyed in a natural cataclysm. His ship landed in the American countryside, near the fictional town of Smallville. He was found and adopted by farmers Jonathan and Martha Kent, who named him Clark Kent. Clark developed various superhuman abilities, such as incredible strength and impervious skin. His adoptive parents advised him to use his abilities for the benefit of humanity, and he decided to fight crime as a vigilante. To protect his privacy, he changes into a colorful costume and uses the alias "Superman" when fighting crime. Clark Kent resides in the fictional American city of Metropolis, where he works as a journalist for the Daily Planet. Superman's supporting characters include his love interest and fellow journalist Lois Lane, Daily Planet photographer Jimmy Olsen and editor-in-chief Perry White. His classic foe is Lex Luthor, who is either a mad scientist or a ruthless businessman, depending on the story."*

Start by processing the text:

```
doc = nlp(text)
```

This performs *sentence splitting* as part of the analysis pipeline. Now add code to print all the resulting sentences.

(b) In addition to the constituent parse trees discussed in the lecture, dependency parsing[7] is another formalism in the family of English grammars. In dependency parse trees, two tokens are connected by a single arc, where the label of the are is the dependency relation. The start node of the arc is known as the *Governor* (*head* in spaCy) and the token at which the arc points at is known as the *Dependant* (*child* in spaCy). A *Dependant/Child* can only have one *Governor/Head* but the vice versa is not true.

---

[7]See https://web.stanford.edu/~jurafsky/slp3/14.pdf and https://spacy.io/usage/linguistic-features#dependency-parse for more details on dependency parse trees

Now, try to use spaCy's "displaCy visualizer"[8] to visualize the dependency parse tree for the sentence given in your worksheet.

---

[8]See https://spacy.io/usage/visualizers