CS 421 – Natural Language Processing – Spring 2019
Term Project: Part 1

Worth: 100 points
Deadline: Mon 4/15, by 11:59pm

This is the first and easier part of the project.
Goal/Tasks: 1) choose one of the parsers listed below, parse all the sentences listed in Section 1.2, and others which are possible given the specifications; 2) develop a module that infers the domain from the category of the question.

1 Task 1: Parsing

You have a choice of parsers you can use, as follows:

1. Stanford parser / CoreNLP (in Java): `https://nlp.stanford.edu/software/lex-parser.html`
2. OpenNLP (in Java): `http://incubator.apache.org/opennlp/`
3. NLTK (in Python): `https://www.nltk.org/`

Notes.

- These parsers all differ as concerns their underlying approach, the kind of grammar they support, etc. It is up to you to choose the one you feel more comfortable with and that provides you with a parse tree you can productively use for the second part of the project.
- Using one of these parsers shortcircuits your work, since you do not have to write the grammar yourself. On the other hand, these parsers may make mistakes, or may return unsatisfactory results.
- You are asked to use the parser on the sentences in Section 1.2, but of course you should not limit yourself to those. They don't represent all the structures that are possible given the specifications.
- You may want to write a simple script or program that systematically does the testing of all your sentences, and writes the resulting trees to file.

1.1 Grammar

You will have to parse only questions (no statements or imperatives), both *yes-no* questions, like *Is Rome the capital of Italy?*, and *wh-questions*, like *Which actress won the oscar in 2012?* (more examples are given later). Note that the proper nouns come from the three databases that this project makes use of. The file sqlite.pdf (to be soon available in the same Blackboard folder) provides some information on how to use SQLITE, at this point, only to access the DBs, not to query them. Accessing the DBs is not actually necessary for Part 1, but in case you want to have a sense of the kind of data we have, you can start having a look.
In general, you can assume that NP and VPs are simplified as follows:

NPs can be:

1. proper nouns: we will simplify people names by only using last names when they exist e.g. *Swift, Kubrik*, or their only name *Madonna*; for *Lady Gaga*, let's use *Gaga*. One problem: among actors, there are two Hepburns (Audrey and Katherine), and two Kelly (Gene and Grace), and perhaps others. Don't worry about this for Part 1, just use the last name;

2. titles of movies, albums or tracks: when they are not NPs, for example *I Miss You*, you can simplify them by creating a new proper noun by concatenating the words *IMissYou*;

3. common nouns such as *continent, capital, border, river, mountain, actor, movie, track, singer, artist, album, rock, pop, dance*, etc;

4. note that *oscar* is not capitalized so it is treated as a common noun, but feel free to experiment with capitalization:

5. nominals such as *mountain chain, rock album*

6. complex NPs with determiners, adjectives and numerals: *the best movie, the highest mountain, the last track*;

7. complex NPs followed by one or two prepositional phrases: *movie by Kubrik with Nicholson*

8. wh-NPs that include a so-called wh-word, i.e., *who, which, when*. Note that *who* is a pronoun, *which* is used here as a determiner, *when* an adverb.

VPs include

1. an NP only, e.g., *win the oscar*;

2. a prepositional phrase only, e.g., *in which continent does Canada lie; did Neeson star in Schindler's List*;

3. an NP and one or two prepositional phrases, *win the oscar in 2012, release a new album in february in Italy*.

1.2 Examples

Some examples were given above, some more are provided here.
IMPORTANT: The examples given below are meant as ... examples! You should not limit yourself to running these questions, think about others that are possible.
Yes-no questions:

(1a) Is Rome the capital of Italy?

(1b) Is France in Europe?

(1c) Is the Pacific deeper than the Atlantic?

(1d) Did Neeson star in Schindler's List?

(1e) Did Swank win the oscar in 2000?

(1f) Is the Shining by Kubrik?

(1g) Did a French actor win the oscar in 2012?

(1h) Did a movie by Spielberg with Neeson win the oscar for best film?

(1i) Did Madonna sing PapaDoNotPreach?

(1j) Does the album Thriller include the track BeatIt?

(1k) Was Beyonce' born in the USA?

Wh-questions:

(2a) Who directed Hugo?

(2b) Which is the scary movie by Kubrik with Nicholson?

(2c) Who won the oscar for the best actor in 2005?

(2d) Which actress won the oscar in 2012?

(2e) Who directed the best movie in 2010?

(2f) In which continent does Canada lie?

(2g) What is the capital of Spain?

(2h) With which countries does France have a border?

(2i) Which is the highest mountain in the world?

(2j) Where is the highest mountain?

(2k) Which is the deepest ocean?

(2l) Which pop artist sings CrazyInLove?

(2m) Where was Gaga born?

(2n) In which album does Aura appear?

(2o) Which album by Swift was released in 2014?

1.2.1 Notes and Simplifying Assumptions

- No morphological processing.
- You can assume there never are more than two adjectives or prepositional phrases in sequence.
- For unknown words (i.e., most proper nouns), the Stanford parser and OpenNLP postulate a Proper Noun when they see a capital letter (see example with "XYZ" in the examples at the end of this file). Instead NLTK fails on unknown words, but it is easy to add new words to NLTK with simple CF productions such as `Shining` $\rightarrow$ `NNP`.
- Numbers such as 2012 are automatically recognized as a number.
- No conjunctions.

2 Task 2: Determine category

You will have to infer the category (Geography, Music, or Movies) from the question itself. This is necessary to build the right SQL query that uses the right database / tables. You are free to use whatever approach you may deem reasonable, including word embeddings or lexical resources such as WordNet (but not querying the web). Note that just indexing from words to category won't work, since eg country names can appear in each category (*Was Madonna born in Italy?*); likewise, we can ask about date of birth for all persons (actors, directors, singers). In fact, some questions are ambiguous category wise, especially music vs movies: whereas we know that the verb phrase *to be born* can only apply to animate entities, hence not to geography, we don't know if we are talking about a singer or about an actor/actress or director.

3 What and how to hand it in

You'll write a couple of paragraphs on why you chose the specific parser you did. You'll include the parse trees the parser returns for the following sentences: (1c), (1e), (1f), (1j); (2a), (2b), (2f), (2h), (2m), (2n), (2o). Of course, for each parse tree, also include the corresponding sentence. Additionally include all the code you wrote (not the parser code you downloaded) , including the "category" code.

The file(s) should include your names and UIN's. One of the two of you will submit the file(s) electronically via the website. Additional details on submission will be sent by email.

4 A few examples

Here are a few examples with their parses from the Stanford parser (from the online demo at `http://nlp.stanford.edu:8080/parser/index.jsp`). The Stanford parser is used only because it has an easy-to-use online demo, not because it is the best of the three parsers.

```
which pop artist sang PapaDoNotPreach?

(ROOT
  (SBARQ
    (WHNP (WDT which)
      (ADJP (JJ pop))
      (NN artist))
    (SQ
      (VP (VBD sang)
        (NP (NNP PapaDoNotPreach))))
    (. ?)))

Who released a rock album in 2012?

(ROOT
  (SBARQ
    (WHNP (WP who))
    (SQ
      (VP (VBD released)
        (NP (DT a) (NN rock) (NN album))
        (PP (IN in)
          (NP (CD 2012)))))))

In which country was Niven born

(ROOT
  (SBARQ
    (WHPP (IN In)
      (WHNP (WDT which) (NN country)))
    (SQ (VBD was)
      (NP (NNP Niven))
      (VP (VBN born)))))
```

```
Did Redmayne win the oscar for best actor in 2015?

(ROOT
  (SQ (VBD Did)
    (NP (NNP Redmayne))
    (VP (VB win)
      (NP
        (NP (DT the) (NN oscar))
        (PP (IN for)
          (NP (JJS best) (NN actor))))
      (PP (IN in)
        (NP (CD 2015))))
    (. ?)))

Who starred in Schindler's List?

(ROOT
  (SBARQ
    (WHNP (WP Who))
    (SQ
      (VP (VBD starred)
        (PP (IN in)
          (NP
            (NP (NNP Schindler) (POS 's))
            (NN List)))))
    (. ?)))

what is the capital of XYZ

(ROOT
  (SBARQ
    (WHNP (WP what))
    (SQ (VBZ is)
      (NP
        (NP (DT the) (NN capital))
        (PP (IN of)
          (NP (NNP XYZ)))))))
```

```
With which countries does Italy have a border

(ROOT
  (PP (IN With)
    (SBAR
      (WHNP (WDT which))
      (S
        (NP (NNS countries))
        (VP (VBZ does)
          (SBAR
            (S
              (NP (NNP Italy))
              (VP (VBP have)
                (NP (DT a) (NN border)))))))))

Is Italy in Europe?

(ROOT
  (SQ (VBZ is)
      (NP (NNP Italy))
      (ADVP (RB in))
      (NP (NNP Europe))
      (. ?)))
```