NLP Midterm: Sean Kennedy, Southern Methodist University ### Q1: ***a. [5 pts] Define homonymy and polysemy and give an example of each.*** ***Homonymy:*** when two or more words have the same spelling or pronunciation but different meaning. *Example **cap**: a cap could be a hat, or a cap as in terms of a limit.* ***Polysemy***: when the same word has multiple meanings. *Example **good**: could be used as an adjective, or as a noun as in terms of the "greater good".* ***b. [5 pts] Define NLU and NLG and give an example of each.*** ***NLU*** *Natural Language Understanding*: A subdivision of **NLP**, this particular study discipline involves comprehension of written text or audio - often in the form of summary extraction or association. ***NLG*** *Natural Language Generation*: The inverse of **NLU**, this subdivision of **NLP** seeks to create text or audio that has the properties of real text. ### Q2: You are given the following grammar for expressions: E → I E → E + E E → E * E E → (E) I → a I → b I → 0 I → 2 ***a. [10 pts] Show parse tree(s) for the expression '2 + 2 * 2'*** **From the start symbol** (1) E → E + E (addition operator) (2) E → E + E * E (multiplication operator) (3) E → 2 + E * E (substitution in position 1) (4) E → 2 + 2 * E (substitution in position 2) (5) E → 2 + 2 * 2 (substitution in position 3) **With middle substitution** (1) E → E + E (addition operator) (2) E → 2 + E (substitution operator) (3) E → 2 + E * E (multiplication operator) (4) E → 2 + 2 * E (substitution in position 2) (5) E → 2 + 2 * 2 (substitution in position 3) ***b. [10 pts] Describe any interesting observations in your answer to a.*** Two separate parse trees could lead to different numerical answers depending on the order in which operators and substitutions are applied. In this particular case, the order of operations in which the actual mathematics is applied makes a difference on the outcome. For instance - interpreting the input to be 2 * 2 (which is 4) and then **adding** 2 - yields a result of **6**. Whereas, taking 2 + 2 (which is also 4) and then **multiplying** by 2 - yields a result of **8**. As such, this particular statement is ambiguous as to what the intended output really is. Using the parentheses operator (), one could enforce the correct associations for the multiplicative and additive functions. ### Q3: Consider the following grammar and sentence: ![Grammar](midterm_grammar.png) *Sentence: I booked a flight from LA* ***a.[10 pts] In what way is this sentence ambiguous? Describe different interpretations of this sentence.*** The sentence could be interpreted in one of two ways - the Proper Noun LA can be thought of as a **place** that you may have booked a flight to in the past or, could be a **person** that you may be booking a flight from (in the case of a travel agent named **LA**). ***b. [10 pts] Show the parse trees for this sentence and where the ambiguity manifests in the parse trees.*** In the parse tree below, we can see that the ambiguity manifests itself in the grouping of the words after the verb **booked**. In the first parse tree, "a flight from LA" is grouped as "a" "flight from LA" which would be the case if I purchsed a ticket from someone named LA. ![ParseTreeOne](midterm_parse_tree_1.png) In the second parse tree, "a flight from LA" is grouped as "a" "flight from LA", splitting the phrase "flight from LA" into smaller Nominal/PP chunks. In this interpretation, I have booked a flight from LA sometime in the past (when the flight occurs or if it has already occurred is unclear) ![ParseTreeOne](midterm_parse_tree_2.png) ### Q4: **The image below shows Google search results for the query “harry potter”** ![HarryPotter](harry_potter.png) **As the results show, the query could represent any of the seven books in the harry potter franchise, any of the film adaptations of the books, a theme park, or a ride, an audiobook, cartoons, et al.** ***a. [10 pts] Discuss why google shows a mix of such results and what factors can influence the search results for this query that will be presented to you.*** The term "harry potter" by itself is very generic given the amount of varied content on the subject. Whenever the search terms are short and generic, it is likely that google will present a generic link (particularly to wikipedia) if there is one available. In this case - the top link is to Wikipedia. If there were a current Harry Potter film in theaters, almost assuredly, the search algorithm would bubble those results to the top. When there are films in the theater, many users will be searching for "new harry potter film" or "harry potter film" or "harry potter and the {insert harry potter film name here}" - when users enter these terms, click the results then navigate to purchase tickets - the algorithm will learn that most users that are generically searching for "harry potter" are most likely looking for tickets. The same would be true if there were a

book that had recently hit the shelves. In this instance - what you are shown is largely driven by the behavior of other users. Another possible set of factors that could affect the results shown to you by google is your own search history - particularly if it is recent or very frequent. Perhaps you are a member of the [Harry Potter Fanclub](wwww.wizardingworld.com), in that case - a search for "harry potter" could very well lead to a link to that website being the top hit. During this time of year, it is entirely possible that folks (particularly children or parents of young children) could be looking for Harry Potter Halloween costumes, which google's algorithms could also infer from historical seasonality and/or the age of the user. ***b. [15 pts] Consider the following sentence:*** *The bank can guarantee deposits will eventually cover future tuition costs because it invests in adjustable-rate mortgage securities.* **The word bank has multiple senses. Use Wordnet to show the top two sense, glossaries and examples for bank and describe (at a high level) how you can use this information to find the proper sense for this word in a sentence.** Wordnet link: http://wordnetweb.princeton.edu/perl/webwn **S: (n) bank (sloping land (especially the slope beside a body of water)) "they pulled the canoe up on the bank"; "he sat on the bank of the river and watched the currents"** **S: (n) depository financial institution, bank, banking concern, banking company (a financial institution that accepts deposits and channels the money into lending activities) "he cashed a check at the bank"; "that bank holds the mortgage on my home"** From the two examples on wordnetweb, we can see the two most likely senses for the word bank are that of a **financial institution** or that of a **slope or grade**. At a high level, simply inspecting the glossaries and examples for words that overlap in the given corpus (in this case a sentence) and counting which one has the most occurrences of overlap would be a great way to determine which sense is the correct sense to use for our interpreter. The first glossary item contains zero words in our corpus whereas the second contains a few words that overlap: **deposits and mortgage**. If we had to choose between one of the two senses, this would be a very simple and effective way to do so. ### Q5: **You are building an online moving streaming service which enables looking up information on movies, genres, directors, actors and customer movie preferences.** ***a. [10 pts] What is the customers intent (i.e. what are they looking for) with the following queries? (these are individual queries, not queries entered in succession)*** *“Drama”, “Jurassic Park”, “Indiana Jones: Raiders of the lost ark”, “Steven Spielberg”* The first query - "Drama" - is clearly someone that is looking for a generic suggestion from a broad genre. A large number of films fit this specific query so a good search algorithm would apply a genre filter to the results and then rank them based on shared rankings from other users. In the second and third queries - "Jurassic Park and Indiana Jones: Raiders Of The Lost Ark" - it is clear that the users are looking for specific films. While the first option - Jurassic Park - is likely to result in more than one possible option (Jurassic Park 1, Jurassic Park 2, 3...10 etc) whereas the second query is a very specific movie from a series of films (the Indiana Jones saga). The final query - "Steven Spielberg" - is looking for films from a specific director - independent of the genre or title. Many of Spielberg's films could fall into the action/adventure or drama categories. This user is a fan of the director, specifically - so search results should be prioritized to only those movies directed by Spielberg (perhaps followed by movies that are highly rated and directed by other directors in the same vein, say, James Cameron). ***b. [5 pts] A customer searches for “Indiana Jones” but clicks on and watches “Jurassic Park” – what insights can you get from this customer action?*** The most likely scenario in this case is that someone is looking for a movie that is similar to something in the "Indiana Jones" tradition - action, adventure, drama, family appropriate. In this case - they are likely to blow past the top search choices (which would most likely be Indiana Jones films) and move directly to the suggested titles that are based on the "Indiana Jones" search string. If a user were to see Jurassic Park on that list, they may see that it was also a Steven Spielberg film and decide that it meets the criteria they were originally hoping to match based on the director. ***c. [10 pts] The customer searches for “Indiana Jones: Raiders of the lost Ark” but it’s not available in their region (US, EU, Asia). What search results would you show the customer? Discuss how you would build that experience from a technical design perspective.*** In this scenario the best suggestion that could be made by the system would involve a

few components: - A matching algorithm to filter results shown to the user based on category / actor / director etc. - A ranking system to sort results based on known user preferences. Are there actors/ directors that the user has preferred in the past? (**Harrison Ford?**) - Another ranking system that ranks movies that the user has not yet seen. But this time based on the mutual ratings that they share with other users in the same domicile - creating a panel of users - and then looking for movies that would be highly rated by the panel. - Combining these systems into one ranking system which is a mixture of textual based matching and user/panel-based rankings so that the user has options to choose from. Most of this could easily be accomplished with python - **NLTK** for text processing and **sklearn** for affinity matrix and user clustering algorithms and the handling of sparce matrices.