Your Name:

Your Andrew ID:

Homework 1

Collaboration and Originality

1. Did you receive help of any kind from anyone in obtaining your data for this assignment (Yes or No)? It is not necessary to describe discussions with the instructor or TA.

   If you answered Yes, provide the name(s) of anyone who provided help, and describe the type of help that you received.

2. Did you give help of any kind to anyone in obtaining their data for this assignment (Yes or No)?

   If you answered Yes, provide the name(s) of anyone that you helped, and describe the type of help that you provided.

3. Are you the author of every word of your report (Yes or No)?

   If you answered No:
   a. identify the text that you did not write,
   b. explain where it came from, and
   c. explain why you used it.

Your Name:

Your Andrew ID:

You must follow this report template. Instructions are given in red italic font throughout the template. Delete the instructions from your report before submitting it.

At a high level, this assignment consists of two parts: i) Do an experiment, and ii) analyze and discuss the experiment. The template includes guidance about what to discuss, but consider them general advice, rather than strict instructions. You may discuss other aspects of the experiment that you find useful or interesting. Grading is based more on the quality of your analysis than a specific checklist.

1 Topic Investigation

1.1 Topics

Topics From Noun Phrases	Supporting Noun Phrases

Topics From (fill in type)	Supporting (fill in type) Entities

1.2 Analysis

Discuss the value of noun phrase and entity distributions as a way of discovering the topics covered by an unknown corpus. Are there interesting differences in the type of information provided by different

sources of information (noun phrases, different types of entities), or in information obtained from different parts of the frequency distribution, or in ctf vs. df frequency distributions? Discuss whether frequency distributions are a useful method of discovering what a corpus contains; justify your answer. Include any other interesting observations that you may have.

2 People Investigation

Identify the three people that this experiment investigates.

2.1 Associated Topics

For each of the three people that you investigated:

- *Provide a brief description that was obtained by examining sample documents; and*
- *Provide a brief description of what new information (if any) was learned by seeing noun-phrases that are strongly associated with that person.*

2.2 Associated People

For each of the three people that you investigated:

- *Provide a list of the co-occurring individuals that you investigated; and*
- *Provide a brief description of what new information (if any) was learned by seeing co-occurring individuals.*

2.3 Associated Organizations

For each of the three people that you investigated:

- *Provide a list of the co-occurring organizations that you investigated; and*
- *Provide a brief description of what new information (if any) was learned by seeing co-occurring organizations.*

2.4 Analysis

Discuss the value of co-occurring noun phrases and entities as a way of discovering more information about people discussed in the news. Are there interesting differences in the type of information provided by different types of co-occurring information (noun phrases, different types of entities), or in information obtained from entities that have different strengths of co-occurrence? Is co-occurrence analysis a useful method of discovering significant information about an individual; justify your answer. Include any other interesting observations that you may have.

Hint: When examining co-occurrences, you should spend a few minutes determining whether there is actually a relationship between the two entities and what it might be (e.g., by doing a web search and examining a few documents). High PMI doesn't necessarily mean that the pair of entities is related in the way that you expect.

3 Organization Investigation

Identify the three organizations that this experiment investigates.

3.1 Associated Topics

For each of the three organizations that you investigated:

- *Provide a brief description that was obtained by examining sample documents; and*
- *Provide a brief description of what new information (if any) was learned by seeing noun-phrases that are strongly associated with that organization.*

3.2 Associated People

For each of the three organizations that you investigated:

- *Provide a list of the co-occurring individuals that you investigated; and*
- *Provide a brief description of what new information (if any) was learned by seeing co-occurring individuals.*

3.3 Associated Organizations

For each of the three organizations that you investigated:

- *Provide a list of the co-occurring organizations that you investigated; and*
- *Provide a brief description of what new information (if any) was learned by seeing co-occurring organizations.*

3.4 Analysis

Discuss the value of co-occurring noun phrases and entities as a way of discovering more information about organizations discussed in the news. Are there interesting differences in the type of information provided by different types of co-occurring information (noun phrases, different types of entities), or in information obtained from entities that have different strengths of co-occurrence? Is co-occurrence analysis a useful method of discovering significant information about an organization; justify your answer. Include any other interesting observations that you may have.

4 Using Google to Calculate PMI and Phi-Square

4.1 Calculation

Entity Pairs	Frequency Counts			PMI	Phi-Square
	E_1	E_2	E_1AND E_2		
Jamie Callan AND Tyler Perry	*00,000,000*	*00,000,000*	*00,000,000*	*0.00000*	*0.00000*
Jamie Callan AND Andrew Moore					
Jamie Callan AND Kevyn Collins-					

Thompson					
K Callan AND Tyler Perry					
Tyler Perry AND Oprah Winfrey					

Sort this table in order of descending PMI.

Analysis

Discuss and provide an analysis of your experimental results. You may wish to consider some of the issues listed below (i.e., these are example issues, not required issues).

- *Do PMI and Phi-Square identify meaningful relationships among pairs of entities? What is the nature of the (real world) relationships that you observed?*
- *Do the metrics behave as you expected? If the metrics make mistakes, what might have caused the behavior that you observed?*
- *Do the metrics tend to agree or disagree? When they disagree, do you understand why?*
- *Do you consider one metric more effective in general, or more effective for high- or low-frequency entities?*

Feel free to discuss other issues of your own choosing. The goal is for you to show that you have thought about and tried to analyze the experimental results that you obtained.

5 People Investigation Using Google

5.1 Calculation

Entity 1

Entity name (E_1):

Frequency (E_1):

Why you picked this entity:

Sort this table in descending order of PMI					
		Frequency			
	Entity E_i	**E_i**	**E_1, E_i**	**PMI**	**Phi-Square**
1					
2					
3					
4					
5					

Sort this table in order of descending PMI.

Entity 2

Entity name (E_2):

Frequency (E_2):

Why you picked this entity:

Sort this table in descending order of PMI					
	Entity E_i	**Frequency**		**PMI**	**Phi-Square**
		E_i	**E_2, E_i**		
1					
2					
3					
4					
5					

Sort this table in order of descending PMI.

Entity 3

Entity name (E_3):

Frequency (E_3):

Why you picked this entity:

Sort this table in descending order of PMI					
	Entity E_i	**Frequency**		**PMI**	**Phi-Square**
		E_i	**E_3, E_i**		
1					
2					
3					
4					
5					

Sort this table in order of descending PMI.

Analysis

This experiment was similar to the first experiment, however there were several significant differences: i) Google vs. Sifaka; ii) current web documents vs. older Wall Street Journal documents; and iii) Phi-Square information in addition to PMI. Discuss and provide an analysis of your experimental results. You may wish to consider some of the issues listed above (i.e., these are example issues, not required

issues). Was there a difference in the amount of work that needed to perform? Which approach do you prefer, and why?