MODEL A: A Poem *by John Godfrey Saxe (1816-1887)*

THE BLIND MEN AND THE ELEPHANT

It was six men of Indostan
To learning much inclined,
Who went to see the Elephant
(Though all of them were blind),
That each by observation
Might satisfy his mind.

The *First* approach'd the Elephant,
And happening to fall
Against his broad and sturdy side,
At once began to bawl:
"God bless me! but the Elephant
Is very like a wall!"

The *Second*, feeling of the tusk,
Cried, -"Ho! what have we here
So very round and smooth and sharp?
To me 'tis mighty clear
This wonder of an Elephant
Is very like a spear!"

The *Third* approached the animal,
And happening to take
The squirming trunk within his hands,
Thus boldly up and spake:
"I see," quoth he, "the Elephant
Is very like a snake!"

The *Fourth* reached out his eager hand,
And felt about the knee.
"What most this wondrous beast is like
Is mighty plain," quoth he,
"'Tis clear enough the Elephant
Is very like a tree!"

The *Fifth*, who chanced to touch the ear,
Said: "E'en the blindest man
Can tell what this resembles most;
Deny the fact who can,
This marvel of an Elephant
Is very like a fan!"

The *Sixth* no sooner had begun
About the beast to grope,
Then, seizing on the swinging tail
That fell within his scope,
"I see," quoth he, "the Elephant
Is very like a rope!"

And so these men of Indostan
Disputed loud and long,
Each in his own opinion
Exceeding stiff and strong,
Though each was partly in the right,
And all were in the wrong!

MORAL:

So oft in theologic wars,
The disputants, I ween,
Rail on in utter ignorance
Of what each other mean,
And prate about an Elephant
Not one of them has seen!

Search I: Text Search

This activity explores how search engines work and how to use them more effectively. Before considering Web-based search engines, we start with simpler searches.

Before starting this activity, read the poem in Model A. This may be done silently as individuals, or you may select a Reader to read the poem to the entire team. Once everyone has read the entire poem, proceed to the next section and **try to answer questions without looking at the poem**.

A. (12 min) Searching Text	start time:

1. What is compared to:

a.	a rope	
b.	a spear	

2. **Manager**, assign one word pair below (a, b, or c) to each team member. (If the team has more than 3 people, double up on word pairs to check your answers.) **Individually**, try to remember if **each word appears** in the poem (yes) or not (no).

Team Member:	a.		b.		c.	
word #1	tail		tiger		desk	
word #2	foot		whale		over	

3. Individually, decide which word was easier to answer, and why.
For example, was it easier for the first member to find “tail” or “foot” in the poem? Why?

a.	tail & foot	
b.	tiger & whale	
c.	desk & over	

4. Does the word "car" appear in the poem in Model A?

5. For many students, answering question 4 is quick and easy, and doesn't require rereading the poem. Explain why this is the case.

INFORMATION - Full Text Search

At a basic level, **full text search** is the process of going through the entire text one word at a time, and deciding whether each word is the right one.

4. Given an unfamiliar text document, decide if a **full text search** is required for each of the following, and explain your reasoning in each case.

		Is full text search needed? Explain.
a.	Find the number of times that the word "moose" appears.	
b.	Find the total number of words.	
c.	Find whether the phrase "everybody loves parfait" appears.	
d.	Find the document name.	

5. If searching 10,000 words takes 0.001 seconds, how long will it take to search:

	# words	**example**	**search time**
a.	10,000	typical novelette	0.001 sec
b.	50,000	typical novel	
c.	50,000,000	typical encyclopedia	
d.	2,500,000,000	English language Wikipedia	

6. Based on these calculations and your experience, is **full text search** the likely approach to search large sets of documents such as Wikipedia or other large websites?

B.(10 min) Saving Search Results	start time:

word	**document**	**position**
beast	Blind Men & The Elephant	line 27, word 5
Indostan	Blind Men & The Elephant	line 1, word 6 line 43, word 6
raven	The Raven	line 44, word 6
tiger	(not found)	(not found)

1. If we had a website with many poems and stories, users would often search for specific words. Explain why the same search might be repeated.

2. Rather than **repeat** a full text search, a computer could **save previous** results as a list of words with the document and position where each is found (see above). A set of items saved for easy reuse is called a **cache**.
Use the word **cache** to explain why **later searches** will be faster than the **first**.

3. Why is it useful to cache searches that have **no results** (such as *tiger*)?

4. Why is it useful to cache a search result **before** anyone searches for the word?
(This is called **pre-caching**.)

5. Why does a cache require **extra storage space**?

6. Explain why this caching is **more complex** (but still useful)
when the text changes over time.

7. The approach explored in the previous questions is called **indexed search**.
In complete sentences, define indexed search and summarize some of its limitations.

C.(5 min) Variations of a Word	start time:

1. In search systems, it can be useful to treat **different** words as if they were the **same**. For each set of words below, **explain**:
   - What the four words have in **common**.
   - Why we might treat **all four words** as the **same word**.

Set A: walk walked walking walks	
Set B: am be is was	
Set C: big giant huge large	

2. Which sets above (A, B, or C) involve:

a.	Variations of a word based on **spelling**? (This is called **stemming**.)	
b.	Variations of a word based on **meaning**? (This is called **lemmatization**.)	

3. Explain which seems easier for a computer, and why:
stemming or **lemmatization**.

4. In complete sentences, explain why a **search system** might benefit from **stemming** and **lemmatization**.

D. (15 min) Which Words are Useful?	start time:

Set A: blind, elephant, Indostan	**Set B:** a, and, is, had, the

1. Consider the **two sets** of words above.

a.	Which set (A or B) are **most frequent**?	
b.	Which set (A or B) would be **most useful** as search terms to find the poem in Model A?	

2. **Describe** and **explain** the relationship between frequency and usefulness.

3. Some words are (nearly) useless for search, and are ignored by a search engine. Refer to earlier questions about most and least useful search terms, and describe some characteristics of these **stop words**.

A. Word	B. TF	C. DF	A. Word	B. TF	C. DF	A. Word	B. TF	C. DF
the	25	14.203	each	4	0.127	all	2	0.575
and	13	6.922	man,men	4	0.264	approach(ed)	2	0.035
is,be,was,were	12	8.085	this	4	1.214	beast	2	0.004
elephant	10	0.004	what	4	0.761	began,begun	2	0.141
a,an	9	6.537	about	3	0.564	can	2	0.659
of	8	6.666	blind(est)	3	0.008	clear	2	0.044
to	8	4.081	he	3	1.875	exceed(ing)	2	0.006
like	7	0.367	quoth	3	0.000	happen(ing)	2	0.118
very	7	0.253	see	3	0.428	Indostan	2	0.000
his	6	1.161	so	3	0.488	most	2	0.159
I,me	5	2.564	who	3	0.656	on	2	1.602
in	5	4.509				that	2	2.211
						them	2	0.437
						though	2	0.098
						within	2	0.074

4. The poem in Model A contains roughly 300 words. In the table above, **column A** contains ~40 words (or related words) that appear **at least twice**.

a.	List 2 examples of **stemming**.	
b.	List 2 examples of **lemmatization**.	
c.	List 5 **stop words**.	
d.	Find 5 **non**-stop words.	

4. **Column B** lists the **number of times** each word appears in the text.
This is the **term frequency (TF)**.
What is the TF for:

a.	and	
b.	blind	
c.	on	

5. **Column C** lists how often the word would appear in a **typical** document of the same length. This is the **document frequency (DF)**.
Which 5-6 words appear **much more often** in this text than in a typical document?

6. In complete sentences, explain why the expression:

$$\frac{\text{term frequency (TF)}}{\text{document frequency (DF)}}$$

is a useful way to measure the importance of a word in a document.
(This is called **term frequency - inverse document frequency (TF-IDF).**)

7. Optional: How could we estimate the **document frequency (DF)** of a set of words?

E. What are the Key Ideas? (10 min)	start time:

Stanza	Key Words	Score1	Score2
1	men indostan inclined elephant blind mind	763	763
2	elephant fall side bawl elephant wall	162	162
3	tusk sharp clear elephant spear	74	74
4	animal trunk hands spake quoth elephant snake	1048	81
5	hand knee beast quoth elephant tree	971	89
6	ear blindest man marvel elephant fan	85	85
7	grope tail scope quoth elephant rope	1000	91
8	indostan long strong wrong	636	636
9	theological disputants ignorance elephant	83	83

1. We can use word frequencies and TF-IDF to find important parts of a document. The poem in Model A has 9 stanzas (sections). Which stanza (1-9):

a.	is labeled as the **moral**?	9
b.	best describes the **result or outcome** of the story?	
c.	best describes the **topic** of the story?	

2. The table above lists some key words in each stanza. Which stanza(s) contain?

a.	elephant	1,2,3,4,5,6,7,9	c.	indostan	
b.	ignorance		d.	quoth	

3. We add the TF-IDF value for each word in the stanza and divide by the word count to get an **average** score for the importance of each stanza, shown above as **Score1**.

a.	Which 3 stanzas have the **lowest** Score1 values?	
b.	Do these seem like the **least important** stanzas in the poem?	
c.	Which 3 stanzas have the **highest** Score1 values?	
d.	Do these seem like the **most important** stanzas in the poem?	
e.	What key word is in these stanzas but not in any other stanzas?	

4. If we remove the key word in question 3e from the average, we get **Score2**.

a.	Which 2 stanzas have the **highest** Score2 values?	
b.	Do these seem like the **most important** stanzas in the poem?	

5. Explain how we could justify removing this word from the calculation.

Part II of this activity explores some of the issues that arise when searching very large sets of documents, such as Wikipedia, or the entire Web.