National Security Agency (NSA)
Internship

Matt Hoerr

October 22, 2015

Background

Google prioritizes their search results using 4 major methods:

- A Boolean keyword search of how many times the keyword (or part of) shows up in a document
- A Boolean keyword search incorporating synonyms and stems of the keyword
- How popular the document is (Page Rank Algorithm)
- Who pays them

School Website's use Boolean keyword searches as well

- They have a much smaller data set they are prioritizing
- Many problems with this

Social Media could be another focus area

- Many tweets and fb posts are short in length and wont always mention specific keywords

About My Project

In a nutshell, the overall goal of my project is to work around Boolean keyword searches and be able to score and prioritize data sets based on keyword(s)

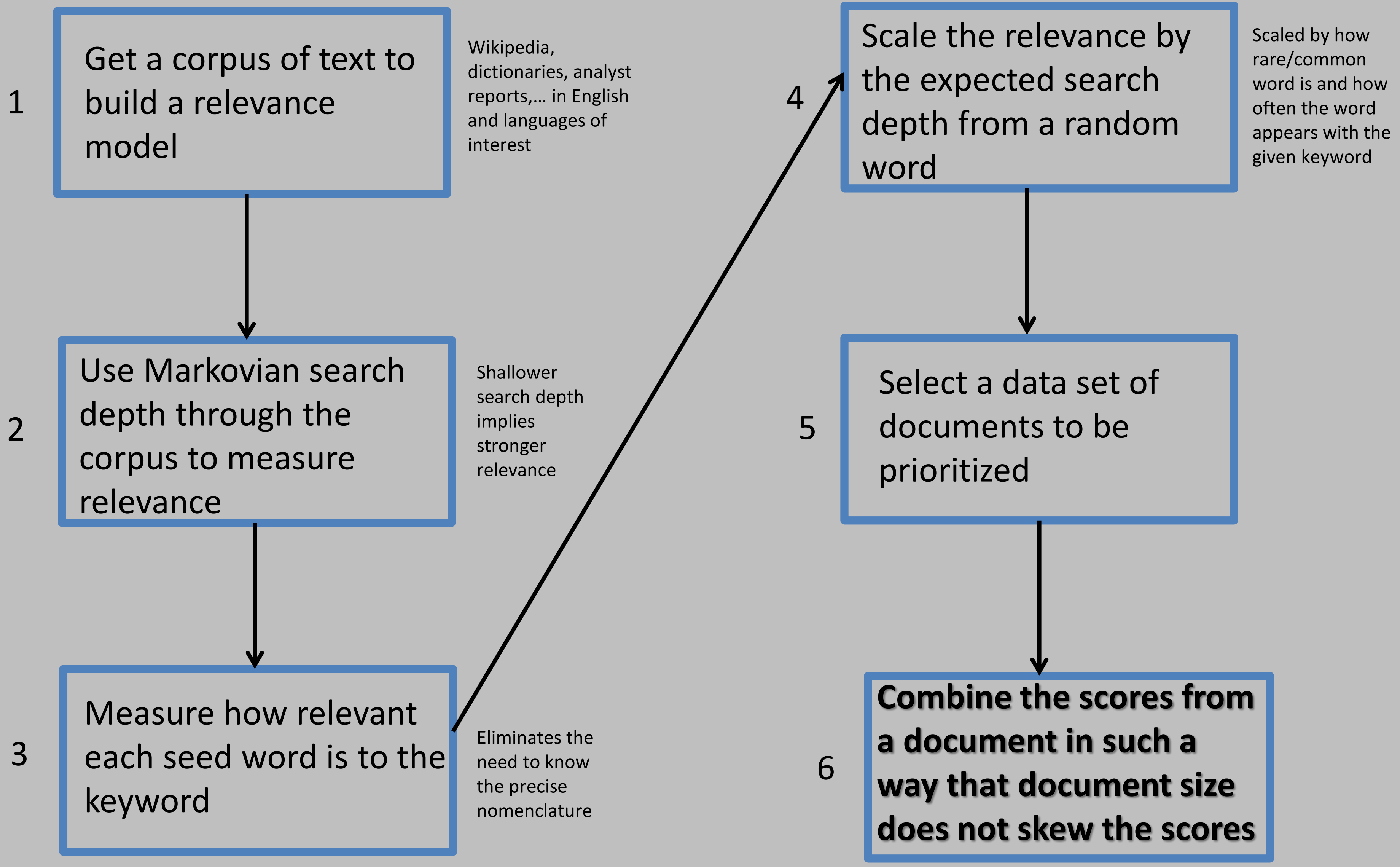
Many documents that may be relevant to a keyword may not contain that keyword, or even a stem/synonym of it

Consider the keyword being ***"Dog"*** and the following 4 sentences being documents you would consider relevant to the keyword

1. The neighbors walked his **dog** while he was on vacation.
2. Her dog is a great guard **dog** and always barks when guests come over.
3. That sheltie loves to fetch tennis balls in the backyard!
4. They say **dogs** are a man's best friend.

The 3rd statement is going to be missed by Boolean keyword searches!!

(U)The Solution



1 Get a corpus of text to build a relevance model

Wikipedia, dictionaries, analyst reports,… in English and languages of interest

2 Use Markovian search depth through the corpus to measure relevance

Shallower search depth implies stronger relevance

3 Measure how relevant each seed word is to the keyword

Eliminates the need to know the precise nomenclature

4 Scale the relevance by the expected search depth from a random word

Scaled by how rare/common word is and how often the word appears with the given keyword

5 Select a data set of documents to be prioritized

6 **Combine the scores from a document in such a way that document size does not skew the scores**

What Could Be Picked Up?

Suppose the keyword is **Europe**

Keyword = Europe

d	σ
Europe	(34861)
European	(248)
Portuguese	(70)
Continent	(22)
Balkan	(12)
Italy	(12)
Italian	(10)
Rome	(7)
⋮	
the	(-.01)
⋮	

Suppose the keyword is **president**

Keyword = president

d	σ
president	(2954)
presidential	(325)
Obama	(239)
Bush	(128)
congressmen	(102)
elected	(65)
veto	(58)
Washington	(51)
debate	(10)
⋮	
the	(.001)
⋮	

Rule of thumb: Any σ over 3 is potentially relevant,
Any σ over 6 is definitely relevant

Stemming

(U)Suppose the keyword is **attack**

d	σ
– **attack**	(2771)
– counter**attack**	(155)
– **attack**ed	(148)
– **attack**ers	(135)

d	σ
– **stab**bed	(24)
– **stab**bings	(18)
– **stab**bing	(16)
– **stab**	(15)

d	σ
– machine**gun**	(84)
– **gun**fire	(63)
– **gun**men	(34)
– **gun**ships	(31)
– **gun**s	(30)
– **gun**	(20)

Important notes here:

1) Recognizes stems of the keyword as relevant

2) Recognizes stems of the relevant words to that keyword as relevant

Scoring and Prioritizing Documents

(U) Each word has a relevance score in association to the given keyword

Mass

The sum of all the word scores in a document

Gives too much weight to larger documents because of the mass accumulation

The more words the more likely the score could be high and not relevant to the keyword

Density

The mean of the scores in a document

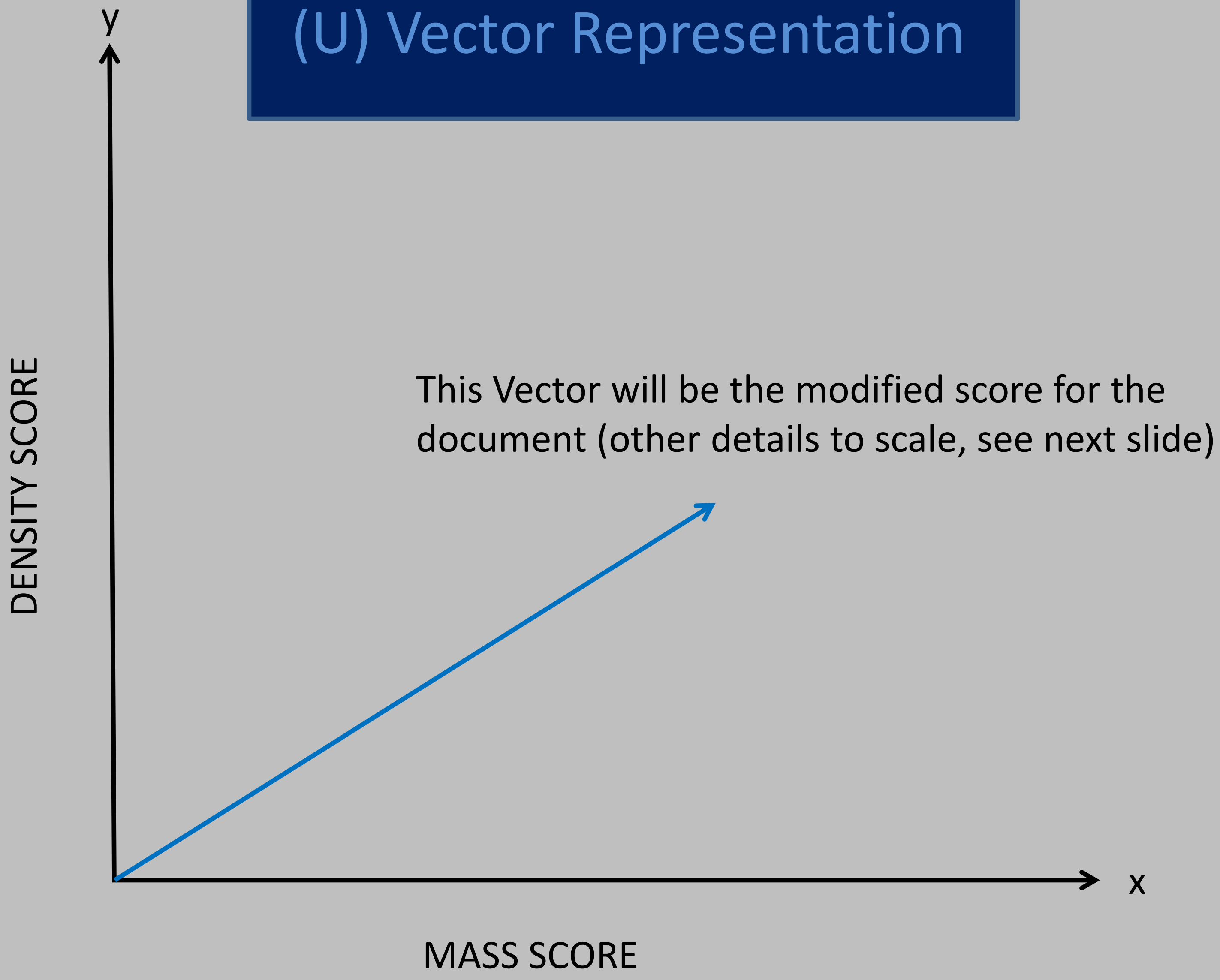
Gives too much weight to smaller documents if they have at least 1 relevant word

If a document is large and relevant, it will still have a score lower than it should because of the large number of words it is dividing by

(U) NOTE: Both methods seem to have opposite strengths and weaknesses

(U) SOLUTION: Take the best of both worlds to create an optimal scoring method

(U) Vector Representation



This Vector will be the modified score for the document (other details to scale, see next slide)

A New Relevance Measure

My Main Contribution

1) If $S_{d,k} \le 1, S_{d,k} = 1$	If the word score ≤ 1, then make it 1
2) $M_{D,k} = \sum_{d \in D} S_{d,k} \ln(S_{d,k})$	Developing Mass Score
3) $D_{D,k} = \left(\frac{1}{\lvert D \rvert} \sum_{d \in D} S_{d,k} \ln(S_{d,k})\right)^2$	Developing Density Score
4) a) $N_R = \sum_{d \in D} \mu(S_{d,k} - 4)$	If word score ≥ 4, consider it relevant
b) $N_{IR} = \sum_{d \in D} \mu(3 - S_{d,k})$	If word score ≤ 3, consider it irrelevant
5) a) $M_{D,k} = M_{D,k}(N_R)^2$	Mass Score *= # relevant words squared
b) $D_{D,k} = D_{D,k}(N_R)^2$	Density Score *= # relevant words squared
c) $M_{D,k} = \frac{M_{D,k}}{N_{IR}}$	Mass Score /= # irrelevant words
d) $D_{D,k} = \frac{D_{D,k}}{N_{IR}}$	Density Score /= # irrelevant words
6) $S_{D,k} = \sqrt{\left(\frac{M_{D,k}}{Mmax - Mmin}\right)^2 + \left(\frac{D_{D,k}}{Dmax - Dmin}\right)^2}$	Developing Final Score

7) $S_{D,k}$ is then put on a Log scale for better interpretation and so 0 acts as a better threshold

Single Keyword Example

Results from using 10,000 open source *Reuters* articles with the keyword ***pipe***:

Ra nk	Article Title	Keyword
1	Chen Stainless Pipe Company Ltd: Key Developments	Yes
2	Yulong Steel Pipe Company Ltd: Key Developments	Yes
3	Russia wants China to prepay for gas to fund pipe	Yes
4	Russian large pipe demand could rise 50%	Yes
5	UPDATE: Pipeline operator Genesis to buy Enterprise's U.S. Gulf business	Yes
6	**U.S. shale oil trade goes waterborne**	**No**
7	**Freepoint expands U.S. natgas, sees oil 'conversion' play**	**No**
8	**Freepoint starts base metals trading, focus on Asia**	**No**
9	Russian pipemakers find silver lining in standoff with West	Yes
10	**CNOOC's Nexen schedules work on Long Lake oil stands upgrader**	**No**

Scoring Documents/Microblogs
Single Keyword

Looking back at our ***"Dog"*** example, lets take a deeper look at the following microblogs:

The	Neighbors	Walked	His	Dog	While	He	Was	On	Vacation
1	1.2	14.3	1	2106	1	1	1	1.14	1.11

→ 5.87

Lets	Go	To	The	Pound	And	Find	Us	A	New	Pet
1.12	1.14	1.14	1	8.7	1	4.1	1	1	2.1	134

→ 2.99

They	Went	To	The	Market	To	Buy	Fruits	And	Vegetables	Today
1	1.14	1.14	1	1.1	1.14	2.3	1	1	1	1

→ -4.24

That	Sheltie	Loves	To	Fetch	Tennis	Balls	In	The	Backyard
1	53	3.1	1.14	118	7.3	21	1	1	16.5

→ 4.47

They	Say	Dogs	Are	A	Mans	Best	Friend
1	1	281	1	1	4.8	4.2	7.3

→ 3.96

Do	You	Want	To	Go	To	The	Zoo	and	Look	At	Zebras
1	1	1.1	1.14	1.14	1.14	1	3.2	1	2.1	1	3.3

→ -2.19

Rule of thumb: 0 is a good threshold to determine relevance,
Anything over 2 is definitely relevant

Multiple Keyword Example

Results from using 10,000 open source *Reuters* articles with the keywords ***pipe & wealth***:

Ra nk	Article Title	Keyword
1	Russian pipemakers find silver lining in standoff with West	Yes
2	**Dubai Investicorp is targeting a more than 30% hike in assets**	**No**
3	Demand for large diameter pipe from Russian energy firms	Yes
4	**Traffic jams, potholes, snail-paced trains, and flights which don't connect are atop the list of frustrations for Russia's businessmen**	**No**
5	**Barratt Developments annual results announcement**	**No**
6	Russian steelmaker Severstal is well placed to benefit from oil pipeline	Yes
7	**Cost cutting is set to remain the main focus for the oil industry**	**No**
8	**U.S. merchant Freepoint Commodities is still open to potential trading asset acquisitions**	**No**
9	**AbTech Oil 2014 milestones: Revenue, Financial, Product Development, and Infrastructure**	**No**
10	Pipelines in the U.S. are undergoing historic realignment	Yes

Scoring Documents/Microblogs
Multiple Keywords

Lets run the keywords ***pipe & wealth*** on the following microblogs:

	The	Pipe	Burst	Will	Cost	The	Company	A	Lot	Of	Money
Pipe	1	2376	102	1	2.1	1	1.6	1	2.2	1	3.6
Wealth	1	2.1	2.6	1	86	1	74	1	1	1	138

→ 14.62

	The	Workers	Worked	Through	The	Heat	To	Finish
Pipe	1	5.1	3.8	1.14	1	2.9	1	1.9
Wealth	1	1.5	1.2	1	1	1.1	1	2.3

→ 0.0

	The	Oil	Spill	Is	Going	To	Cause	Gas	Prices	To	Inflate
Pipe	1	112	41	1	1.16	1	2.1	72	2.6	1	4.2
Wealth	1	18	1.9	1	1	1	2.4	15	89	1	48

→ 23.87

	The	Pipeline	Budget	Will	Be	Done	Very	Soon
Pipe	1	101	3.1	1	1	1.2	1	1.3
Wealth	1	3.1	30	1	1	1.5	1.1	2.1

→ .16

	The	Financial	Report	Will	Be	In	Tomorrow
Pipe	1	1.3	1.1	1	1	1	1.4
Wealth	1	53	9.3	1	1	1	1.3

→ 0.0

Multiple Languages

The goal of multiple languages is to be able to prioritize documents of various languages relative to an English keyword.

Suppose there was a document in your data set was in Spanish and contained the following:

Los vecinos caminaron su perro mientras que él estaba el vacaciones.

(The neighbors walked his dog while he was on vacation.)

By searching the keyword “dog” or something else similar, this document would be flagged as relevant even if it was surrounded by English documents

Challenges of Work Environment

- First time in a real world job
- Developing a pathway for my project on my own
- Learning new languages such as python
- Learning the ways of the company, it is like a whole new way of life

Classroom Experience

- I used many languages I learned from classes here
- I used Linux all summer and my brief knowledge I gained here went a long way
- Projects in the classroom are kind of a small scale example of the real world

What Did I Gain From This?

- Furthered my knowledge of both Java and C++
- Further expanded my knowledge of Linux
- Learned how to write scripts in Python
- Learned how to write PDF code (LaTex)
  - Used this to “code” my scientific paper
- Got to work with very large data sets
- Gained a wealth of opinions and information from coworkers both older and younger
- Got to experience the work world and what it will be like after college
- Made connections to people that I will always have in the future

Questions?