

Surprising Discoveries for Online Health Information

Shweta, Lohitaksh, Neetu, Nalin , Abishek

Summary

- ▶ Task description
- ▶ Techniques you used
- ▶ Results
- ▶ Interpretation of your results
- ▶ Challenge you have had along the way
- ▶ Lessons learnt

Task Description

- ▶ Type of surprise chosen: Personal surprise
- ▶ Steps followed to achieve a surprise:
 - Define expectation → Compute divergence → Compute likelihood → Arrive to a surprise

Techniques used towards surprise:

The basic idea that we have formulated to take out the element of surprise is as follows:-

- ▶ Read the corpus.
- ▶ Clean and concise by removing stop words.
- ▶ Restrict the frequency of words with word length minimum 4 and maximum 20.
- ▶ We run the frequency count of all the filtered words.
- ▶ Run the **TF-IDF** to balance the frequency count and importance-The TF-IDF value of a word increases as the frequency of the word increases which shows that the words appear more frequently in general.
- ▶ Run LSI, Wordcloud and other algorithms

Algorithms used

- ▶ Entropy
- ▶ Mutual Information
- ▶ TFIDF - LSI
- ▶ Word Cloud
- ▶ Cosine Similarity
- ▶ KL divergence (in progress..)

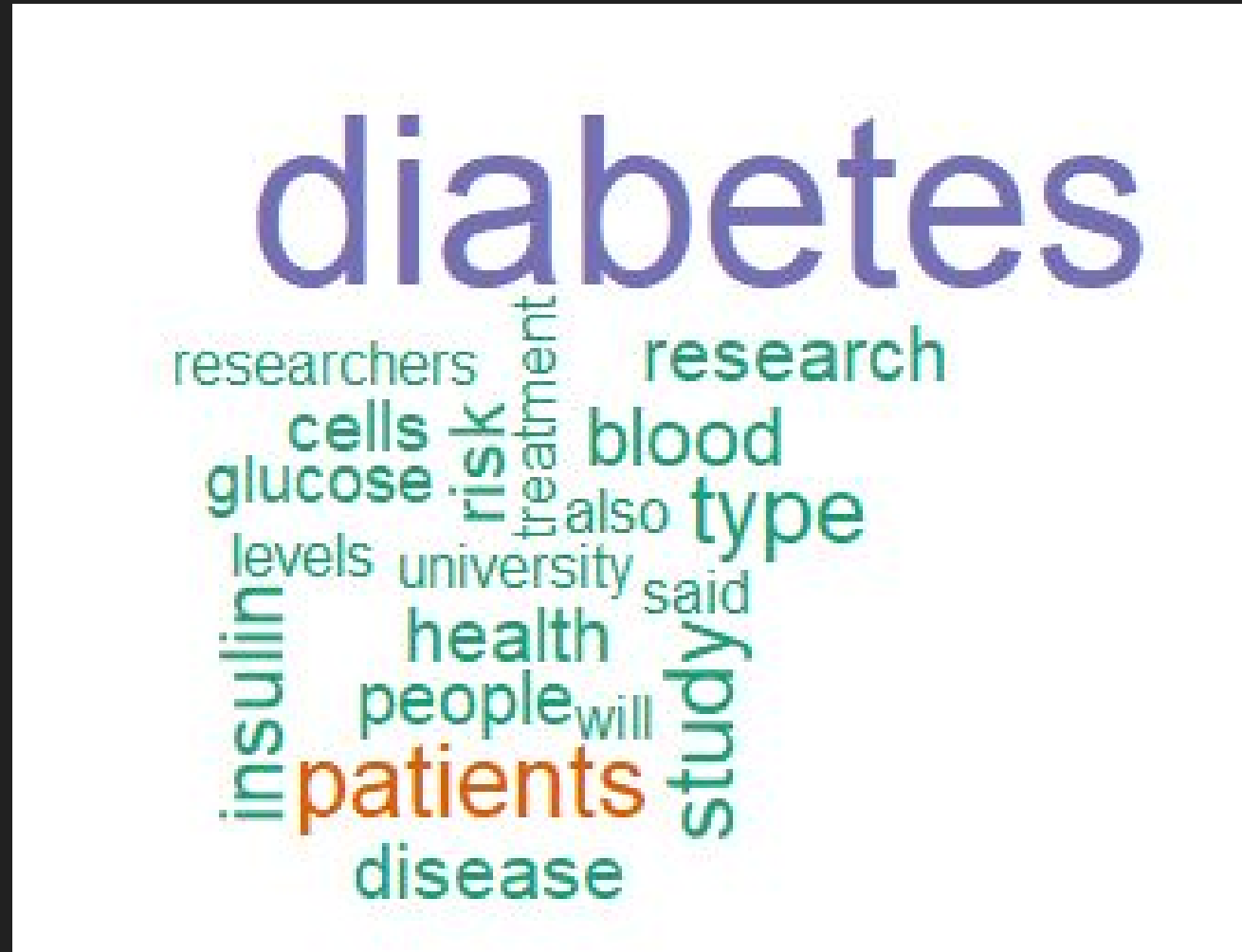
- We use the Kullback Leibler (KL) divergence as the divergence measure since it is a common way to evaluate the divergence between two probability distributions. The surprise score is calculated as Equation 3, where p_{α} is the distribution of the latent themes in an article, q is the distribution of a typical article, and i is the index of the latent themes. We label this approach as KL.

$$s_2 = \text{KL}(p_{\alpha}, q) = \sum_{i=1}^k p_{\alpha i} \log_2 \frac{p_{\alpha i}}{q_i} \quad (3)$$

WordCloud

- ▶ A **word cloud** is a visualization of word frequency in a given text as a weighted list. The technique has recently been popularly used to visualize the topical content of political speeches.
- ▶ In our hypothesis we are trying to see which words are appearing more frequently from the corpus in the wordcloud
- ▶ The larger the word in the cloud the more common the word was in the document

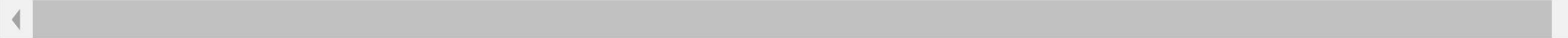
Word Cloud: Forming a Hypothesis



TF - IDF

```
In [5]: from gensim import models
        # train the model
        tfidf = models.TfidfModel(bow_corpus)
        # transform the "system minors" string
        #top 10 frequen
        print(tfidf[dictionary.doc2bow("diabetes patients type study insulin blood risk disease health research".lower().split())])

        print(tfidf[dictionary.doc2bow("come opportunity supporting publication involving millions rapidly suggesting ".lower().split())])
```



```
[(53, 0.3492563305587527), (60, 0.22507556354736905), (62, 0.3512624990218378), (212, 0.30059628126992954), (394, 0.31066264090
650114), (405, 0.3248702030107599), (431, 0.316902484435025), (557, 0.3622871041934056), (636, 0.3007640397348136), (712, 0.299
04426990547595)]
[(204, 0.34994353712370985), (882, 0.34918760631019247), (3230, 0.3528768841850399), (3346, 0.34994353712370985), (4171, 0.3532
8037345900437), (4217, 0.3648741288092756), (4905, 0.358821439314328), (6213, 0.34918760631019247)]
```


Entropy and Mutual Information

- ▶ **Entropy** means the measurement of the change.
- ▶ Bigger is the entropy, more is the word unpredictable.
- ▶ Now our hypothesis is based upon this concept because the higher the randomness of the word it can signify a **surprise** element in such a case.
- ▶ **Mutual Information** works very similar to the decision tree logic. Decision trees use entropy to measure purity of data, mutual information uses entropy to measure information content in data: higher entropy implies lower information.

KL divergence

- ▶ **Kullback-Leibler divergence** (also called relative entropy) is a measure of how one probability distribution diverges from a second, expected probability distribution.
- ▶ Applications include characterizing the relative (Shannon) entropy in information systems, randomness in continuous time-series, and information gain when comparing statistical models of inference.
- ▶ In our hypothesis we are checking how KL divergence between two data matrix and measuring the distance between them
- ▶ The more distance is between the data matrix the more **surprising** it will be.

LSI

- ▶ A key feature of LSI is its ability to extract the conceptual content of a body of text by establishing associations between those terms that occur in similar contexts. - Single Value Decomposition(SVD)
- ▶ SVD is useful - reduced dimensional representation of our matrix that emphasizes the strongest relationships and throws away the noise. In other words, it makes the best possible reconstruction of the matrix with the least possible information.
- ▶ We used TF-IDF score to rank words by their importance and use them to form topics using LSI

Computational Surprise based on LSI

```
2017-12-06 22:39:17,972 : INFO : topic #0(1.350): 0.444*"university" + 0.420*"monash" + 0.273*"health" + 0.230*"32m" + 0.225*"research" + 0.212*"medical" +
0.178*"title" + 0.160*"awarded" + 0.153*"project" + 0.150*"funding"
2017-12-06 22:39:17,978 : INFO : topic #1(1.138): 0.414*"university" + -0.244*"research" + 0.221*"monash" + -0.209*"federal" + -0.166*"health" + -0.161*"medical" +
-0.156*"professor" + -0.156*"cornish" + -0.156*"said" + -0.156*"government"
2017-12-06 22:39:17,984 : INFO : topic #2(1.074): -0.405*"disease" + -0.284*"testicular" + -0.284*"491250" + -0.253*"development" + -0.217*"explore" +
-0.217*"novel" + -0.217*"strategy" + -0.217*"cardiovascular" + -0.217*"312500" + -0.207*"treatment"
2017-12-06 22:39:17,987 : INFO : topic #3(1.047): -0.323*"brain" + -0.207*"traumatic" + -0.207*"erythropoietin" + -0.207*"epo" + -0.207*"injury" + -0.207*"18" +
-0.192*"oxygen" + -0.192*"impact" + -0.192*"556500" + -0.192*"low"
2017-12-06 22:39:17,990 : INFO : topic #4(1.045): 0.409*"health" + 0.220*"title" + 0.197*"medical" + 0.190*"roxon" + 0.190*"minister" + 0.190*"announced" +
0.190*"nicola" + 0.190*"grant" + -0.176*"university" + -0.148*"cornish"
2017-12-06 22:39:17,991 : INFO : topic #0(1.350): 0.444*"university" + 0.420*"monash" + 0.273*"health" + 0.230*"32m" + 0.225*"research" + 0.212*"medical" +
0.178*"title" + 0.160*"awarded" + 0.153*"project" + 0.150*"funding"
2017-12-06 22:39:17,992 : INFO : topic #1(1.138): 0.414*"university" + -0.244*"research" + 0.221*"monash" + -0.209*"federal" + -0.166*"health" + -0.161*"medical" +
-0.156*"professor" + -0.156*"cornish" + -0.156*"said" + -0.156*"government"
```

```
+ -0.167*"end" + -0.153*"type" + 0.138*"ten" + 0.138*"nine" + 0.138*"nondiabetic" + 0.138*"figure" + 0.
131*"pressure"
```

Surprise Element 1

\$673,000 funded to follow up an initial investigation of estrogen as an effective treatment for schizophrenia , allowing a clinical trial in 180 postmenopausal women with schizophrenia using a selective estrogen receptor modulator.

In healthy people, the liver produces glucose during fasting to maintain normal levels of cell energy production. After people eat, the pancreas releases insulin, the hormone responsible for glucose absorption. Once insulin is released, the liver should turn down or turn off its glucose production, but in people with type 2 diabetes, the liver fails to sense insulin and continues to make glucose. The condition, known as insulin resistance , is caused by a glitch in the communication between liver and pancreas.

Metformin, introduced as frontline therapy for uncomplicated type 2 diabetes in the 1950s, up until now was believed to work by making the liver more sensitive to insulin. The Hopkins study shows, however, that metformin bypasses the stumbling block in communication and works directly in the liver cells.

"Rather than an interpreter of insulin-liver communication, metformin takes over as the messenger itself," says senior investigator Fred Wondisford, M.D., who heads the metabolism division at Hopkins Children's. "Metformin actually mimics the action of CBP, the critical signaling protein involved in the communication between the liver and the pancreas that's necessary for maintaining glucose production by the liver and its suppression by insulin."

To test their hypothesis, researchers induced insulin resistance in mice by feeding them a high-fat diet over several months. Mice on high-fat diets developed insulin resistance, and their high blood glucose levels did not drop to normal after eating. Once treated with metformin, however, CBP was activated to the levels of nondiabetic mice, and their blood glucose levels returned to normal. However, when given to diabetic mice with defective copies of CBP, metformin had no effect on blood glucose levels, a proof that metformin works through CBP.

Challenges faced

- ▶ Tools and Machine constraints
- ▶ Large size of the corpus
- ▶ Pre-processing requirements for every algorithm
- ▶ Several algorithms gave us outputs which were not “SURPRISES”, but known facts.

THANK YOU !