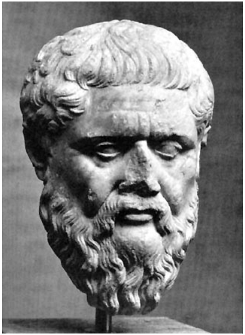
Latent Semantic Analysis Using Google 4-grams

Adam Kapelner*

(Special Thanks to: Dean Foster*)

*Department of Statistics
The Wharton School, University of Pennsylvania

Problem of Knowledge Acquisition



We know much more than we should given our limited experiences.
Plato's spiritual answer: we were born with the knowledge and need "help" revealing it.

Vocabulary $\subset$ Knowledge

We focus on the problem of acquiring vocabulary which we proxy for acquiring knowledge.

Vocabulary $\subset$ Knowledge

We focus on the problem of acquiring vocabulary which we proxy for acquiring knowledge.

Ludwig Wittgenstein wrote an interesting idea about vocabulary in 1953 in his Philosophical Investigations. Summed up,

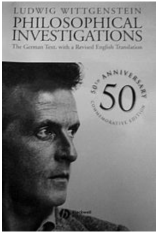
> *"meaning" is "use"*

Vocabulary ⊂ Knowledge

We focus on the problem of acquiring vocabulary which we proxy for acquiring knowledge.

Ludwig Wittgenstein wrote an interesting idea about vocabulary in 1953 in his Philosophical Investigations. Summed up,

> *"meaning" is "use" – words are not defined by reference to the objects they designate, nor by the mental representations one might associate with them, but by how they are used.*

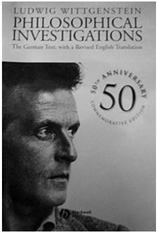

LUDWIG WITTGENSTEIN
PHILOSOPHICAL INVESTIGATIONS
The German Text, with a Revised English Translation
50th ANNIVERSARY
50
COMMEMORATIVE EDITION

Vocabulary ⊂ Knowledge

We focus on the problem of acquiring vocabulary which we proxy for acquiring knowledge.

Ludwig Wittgenstein wrote an interesting idea about vocabulary in 1953 in his Philosophical Investigations. Summed up,

> *"meaning" is "use" – words are not defined by reference to the objects they designate, nor by the mental representations one might associate with them, but by how they are used.*


LUDWIG WITTGENSTEIN
PHILOSOPHICAL INVESTIGATIONS
The German Text, with a Revised English Translation
50th ANNIVERSARY
50
COMMEMORATIVE EDITION

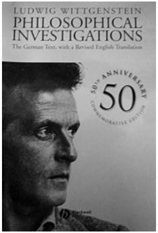
Plato would not be happy that words are not reflective of higher truths...

Vocabulary ⊂ Knowledge

We focus on the problem of acquiring vocabulary which we proxy for acquiring knowledge.

Ludwig Wittgenstein wrote an interesting idea about vocabulary in 1953 in his Philosophical Investigations. Summed up,

> *"meaning" is "use" – words are not defined by reference to the objects they designate, nor by the mental representations one might associate with them, but by how they are used.*


LUDWIG WITTGENSTEIN
PHILOSOPHICAL INVESTIGATIONS
The German Text, with a Revised English Translation
50th ANNIVERSARY
50
COMMEMORATIVE EDITION
Blackwell

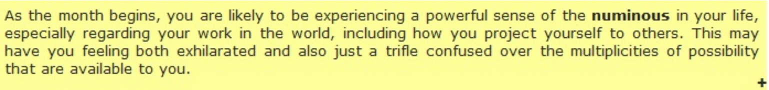
Plato would not be happy that words are not reflective of higher truths... but Plato couldn't explain that...

Vocabulary Acquisition

Children learn 7-15 words per day. Direct instruction nor reading the dictionary can explain this growth. How could it be possible?

Vocabulary Acquisition

Children learn 7-15 words per day. Direct instruction nor reading the dictionary can explain this growth. How could it be possible?

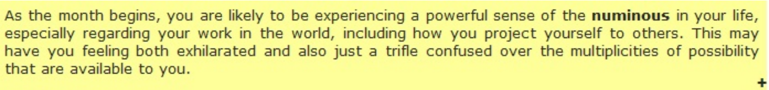


As the month begins, you are likely to be experiencing a powerful sense of the **numinous** in your life, especially regarding your work in the world, including how you project yourself to others. This may have you feeling both exhilarated and also just a trifle confused over the multiplicities of possibility that are available to you.

Answer: Incidental learning from context.

Vocabulary Acquisition

Children learn 7-15 words per day. Direct instruction nor reading the dictionary can explain this growth. How could it be possible?



As the month begins, you are likely to be experiencing a powerful sense of the **numinous** in your life, especially regarding your work in the world, including how you project yourself to others. This may have you feeling both exhilarated and also just a trifle confused over the multiplicities of possibility that are available to you.

Answer: Incidental learning from context.
But how do we learn so well from context?

Latent Semantic Analysis (LSA)

A theory proposed in 1997 by Landauer and Dumais

TK Landauer and ST Dumais. A solution to Platos problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. Psychological review, 1 (2):211240, 1997.

> *"a high-dimensional linear associative model that embodies no human knowledge beyond its general learning mechanism ... domains of knowledge contain vast numbers of weak interrelations that, if properly exploited, can greatly amplify learning by a process of inference"*

Latent Semantic Analysis (LSA)

A theory proposed in 1997 by Landauer and Dumais

TK Landauer and ST Dumais. A solution to Platos problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. Psychological review, 1 (2):211240, 1997.

> *"a high-dimensional linear associative model that embodies no human knowledge beyond its general learning mechanism ... domains of knowledge contain vast numbers of weak interrelations that, if properly exploited, can greatly amplify learning by a process of inference"*

- "latent" — happening behind the scenes

Latent Semantic Analysis (LSA)

A theory proposed in 1997 by Landauer and Dumais

TK Landauer and ST Dumais. A solution to Platos problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. Psychological review, 1 (2):211240, 1997.

> *"a high-dimensional linear associative model that embodies no human knowledge beyond its general learning mechanism ... domains of knowledge contain vast numbers of weak interrelations that, if properly exploited, can greatly amplify learning by a process of inference"*

- "latent" — happening behind the scenes
- "semantic" — based on written language

Latent Semantic Analysis (LSA)

A theory proposed in 1997 by Landauer and Dumais

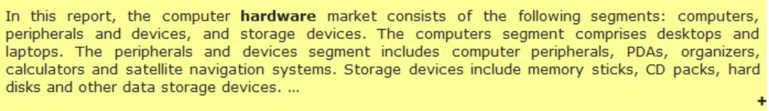
TK Landauer and ST Dumais. A solution to Platos problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. Psychological review, 1 (2):211240, 1997.

> *"a high-dimensional linear associative model that embodies no human knowledge beyond its general learning mechanism ... domains of knowledge contain vast numbers of weak interrelations that, if properly exploited, can greatly amplify learning by a process of inference"*

- "latent" — happening behind the scenes
- "semantic" — based on written language
- "analysis" — an internal computation is made

A Toy Example

Let's say you read the following snippet:

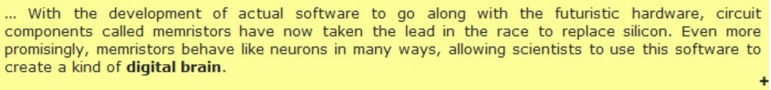


In this report, the computer **hardware** market consists of the following segments: computers, peripherals and devices, and storage devices. The computers segment comprises desktops and laptops. The peripherals and devices segment includes computer peripherals, PDAs, organizers, calculators and satellite navigation systems. Storage devices include memory sticks, CD packs, hard disks and other data storage devices. ...

+

You may get the idea that "hardware" is associated with the word "computer" and this would help you learn the word. As you read more and more such snippets "hardware" becomes more and more associated with "computer"

Now let's read the following:

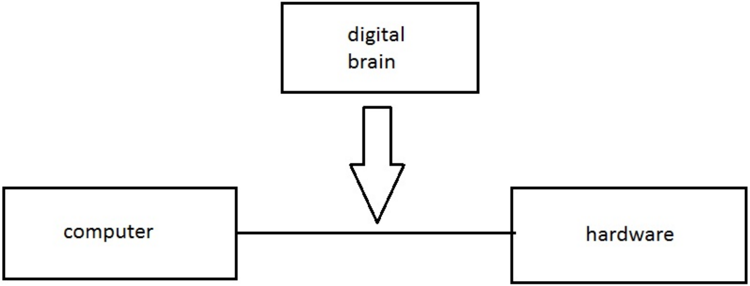


... With the development of actual software to go along with the futuristic hardware, circuit components called memristors have now taken the lead in the race to replace silicon. Even more promisingly, memristors behave like neurons in many ways, allowing scientists to use this software to create a kind of **digital brain**.

+

Dimension Reduction is Key

With dimension reduction the concept of a “digital brain” can become associated with “computer” and “hardware” by “smushing” the “space of knowledge”.



Basic Setup I

How did Landuaer and Dumais lend credibility to their theory?
They ran an experiment.

Basic Setup I

How did Landuaer and Dumais lend credibility to their theory? They ran an experiment.

They took $p \approx 30,000$ encyclopedia articles from Grolier's Academic Encyclopedia, 1980 (truncated at 2000 characters) yielded $n \approx 60,000$ unique words which appeared in more than one article. Call the matrix of word counts per article: $X_{\text{pre}} \in \mathbb{N}_0^{n \times p}$.

Basic Setup I

How did Landuaer and Dumais lend credibility to their theory? They ran an experiment.

They took $p \approx 30,000$ encyclopedia articles from Grolier's Academic Encyclopedia, 1980 (truncated at 2000 characters) yielded $n \approx 60,000$ unique words which appeared in more than one article. Call the matrix of word counts per article: $X_{\text{pre}} \in \mathbb{N}_0^{n \times p}$. Then they made two adjustments, *both* of which were grounded in the cognitive science principles of the time:

1. Each count was converted to log frequencies (adjusting for zeros). Logs mimic the growth of simple learning (empirically documented):

$$x'_{ij} = \ln\left(1 + x_{ij}\right)$$

Basic Setup II

2. Division by entropy (well, the empirical estimate of entropy).

$$\begin{aligned} x''_{ij} &= \frac{x'_{ij}}{\text{entropy}(\mathbf{x}_{i\cdot})} = \frac{x'_{ij}}{-\sum_{l=1}^{p} \mathbb{P}\left(x_{il}\right) \log_2\left(\mathbb{P}\left(x_{il}\right)\right)} \\ &= \frac{\ln\left(1+x_{ij}\right)}{-\sum_{l=1}^{p} \frac{x_{il}}{\sum_{k=1}^{p} x_{ik}} \log_2\left(\frac{x_{il}}{\sum_{k=1}^{p} x_{ik}}\right)} \end{aligned}$$

Basic Setup II

2. Division by entropy (well, the empirical estimate of entropy).

$$\begin{aligned} x''_{ij} &= \frac{x'_{ij}}{\text{entropy}(\mathbf{x}_{i\cdot})} = \frac{x'_{ij}}{-\sum_{l=1}^{p} \mathbb{P}\left(x_{il}\right) \log_2\left(\mathbb{P}\left(x_{il}\right)\right)} \\ &= \frac{\ln\left(1 + x_{ij}\right)}{-\sum_{l=1}^{p} \frac{x_{il}}{\sum_{k=1}^{p} x_{ik}} \log_2\left(\frac{x_{il}}{\sum_{k=1}^{p} x_{ik}}\right)} \end{aligned}$$

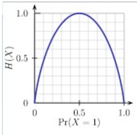
What does this accomplish?

Basic Setup II

2. Division by entropy (well, the empirical estimate of entropy).

$$\begin{aligned} x''_{ij} &= \frac{x'_{ij}}{\text{entropy}(\mathbf{x}_{i\cdot})} = \frac{x'_{ij}}{-\sum_{l=1}^{p} \mathbb{P}\left(x_{il}\right) \log_2\left(\mathbb{P}\left(x_{il}\right)\right)} \\ &= \frac{\ln\left(1 + x_{ij}\right)}{-\sum_{l=1}^{p} \frac{x_{il}}{\sum_{k=1}^{p} x_{ik}} \log_2\left(\frac{x_{il}}{\sum_{k=1}^{p} x_{ik}}\right)} \end{aligned}$$

What does this accomplish? This attenuates the signal of entries with high entropy that is they appear everywhere with high unpredictability in order to amplify the signal of words that appear only in a few places.

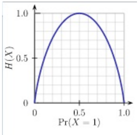


Basic Setup II

2. Division by entropy (well, the empirical estimate of entropy).

$$
\begin{aligned}
x''_{ij} &= \frac{x'_{ij}}{\text{entropy}(\mathbf{x}_{i\cdot})} = \frac{x'_{ij}}{-\sum_{l=1}^{p} \mathbb{P}\left(x_{il}\right) \log_2\left(\mathbb{P}\left(x_{il}\right)\right)} \\
&= \frac{\ln\left(1 + x_{ij}\right)}{-\sum_{l=1}^{p} \frac{x_{il}}{\sum_{k=1}^{p} x_{ik}} \log_2\left(\frac{x_{il}}{\sum_{k=1}^{p} x_{ik}}\right)}
\end{aligned}
$$

What does this accomplish? This attenuates the signal of entries with high entropy that is they appear everywhere with high unpredictability in order to amplify the signal of words that appear only in a few places.



So we are left with:

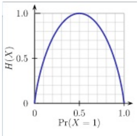
$$X = \text{trans}\left(X_{\text{pre}}\right) \in \mathbb{R}^{n \times p}$$

Basic Setup II

2. Division by entropy (well, the empirical estimate of entropy).

$$
\begin{aligned}
x''_{ij} &= \frac{x'_{ij}}{\text{entropy}(\mathbf{x}_{i\cdot})} = \frac{x'_{ij}}{-\sum_{l=1}^{p} \mathbb{P}\left(x_{il}\right) \log_2\left(\mathbb{P}\left(x_{il}\right)\right)} \\
&= \frac{\ln\left(1+x_{ij}\right)}{-\sum_{l=1}^{p} \frac{x_{il}}{\sum_{k=1}^{p} x_{ik}} \log_2\left(\frac{x_{il}}{\sum_{k=1}^{p} x_{ik}}\right)}
\end{aligned}
$$

What does this accomplish? This attenuates the signal of entries with high entropy that is they appear everywhere with high unpredictability in order to amplify the signal of words that appear only in a few places.



So we are left with:

$$X = \text{trans}\left(X_{\text{pre}}\right) \in \mathbb{R}^{n \times p}$$

Basic Setup III

To exploit interrelationships in this space of knowledge, they run thin SVD to elucidate the structure of the space:

$$X = F_1 W F_2^T \quad \text{where} \quad \underbrace{F_1 \in \mathbb{R}^{n \times p}}_{\text{hanger}}, \quad \underbrace{W \in \mathbb{R}^{p \times p} \geq 0 \ \ \text{diag}}_{\text{stretcher}}, \quad \text{and} \quad \underbrace{F_2^T \in \mathbb{R}^{p \times m}}_{\text{aligner}}$$

Basic Setup III

To exploit interrelationships in this space of knowledge, they run thin SVD to elucidate the structure of the space:

$$X = F_1 W F_2^T \quad \text{where} \quad \underbrace{F_1 \in \mathbb{R}^{n \times p}}_{\text{hanger}}, \quad \underbrace{W \in \mathbb{R}^{p \times p} \geq 0 \;\; \text{diag}}_{\text{stretcher}}, \quad \text{and} \quad \underbrace{F_2^T \in \mathbb{R}^{p \times m}}_{\text{aligner}}$$

where $w_{11} \geq w_{22} \geq \ldots \geq w_{nn}$. Now, shrink the dimensionality of the data by choosing an arbitrary cutoff d, zero out the weights w_{ii} where $i > d$, then "rehydrate" the X matrix using the SVD multiplication:

Basic Setup III

To exploit interrelationships in this space of knowledge, they run thin SVD to elucidate the structure of the space:

$$X = F_1 W F_2^T \quad \text{where} \quad \underbrace{F_1 \in \mathbb{R}^{n\times p}}_{\text{hanger}}, \quad \underbrace{W \in \mathbb{R}^{p\times p} \geq 0 \ \ \text{diag}}_{\text{stretcher}}, \quad \text{and} \quad \underbrace{F_2^T \in \mathbb{R}^{p\times m}}_{\text{aligner}}$$

where $w_{11} \geq w_{22} \geq \ldots \geq w_{nn}$. Now, shrink the dimensionality of the data by choosing an arbitrary cutoff d, zero out the weights w_{ii} where $i > d$, then "rehydrate" the X matrix using the SVD multiplication:

$$X = F_1 \begin{bmatrix} w_{11} & 0 & \\ 0 & w_{22} & \\ & & \ddots \end{bmatrix} F_2^T, \quad \hat{X}^{(1)} = F_1 \begin{bmatrix} w_{11} & 0 & \\ 0 & 0 & \\ & & \ddots \end{bmatrix} F_2^T$$

We call this the "least squares estimate" and denote it $\hat{X}^{(d)}$. It is also the best estimate of X based on the Eckart-Young Thm (lowest Frobenius norm with shrunk rank $d < p$).

Assessment I

ETS TOEFL

They then used 80 retired synonym questions which looked like the following:

```
5. zenith
 a. completion
 b. pinnacle*
 c. outset
 d. decline
```

Simple target word with four alternatives.

Assessment II

For each dimension reduction $\hat{X}^{(d)}$, we can assess the degree of vocabulary learning by having the algorithm "take the test". How?

By computing distances between the target word and each alternative. The distance metric decided upon was the angle in order to ignore raw magnitude effects:

$$\begin{aligned} g^* &= \underset{g \in \{a,b,c,d\}}{\arg\min} \left\{ \theta_{\hat{\mathbf{x}}_{\mathrm{T}\cdot}^{(d)},\ \hat{\mathbf{x}}_{\mathrm{g}\cdot}^{(d)}} \right\} = \underset{g \in \{a,b,c,d\}}{\arg\max} \left\{ \cos\left(\theta_{\hat{\mathbf{x}}_{\mathrm{T}\cdot}^{(d)},\ \hat{\mathbf{x}}_{\mathrm{g}\cdot}^{(d)}} \right) \right\} \\ &= \underset{g \in \{a,b,c,d\}}{\arg\max} \left\{ \frac{\left\langle \hat{\mathbf{x}}_{\mathrm{T}\cdot}^{(d)},\ \hat{\mathbf{x}}_{\mathrm{g}\cdot}^{(d)} \right\rangle}{\left\| \hat{\mathbf{x}}_{\mathrm{T}\cdot}^{(d)} \right\| \left\| \hat{\mathbf{x}}_{\mathrm{g}\cdot}^{(d)} \right\|} \right\} \end{aligned}$$

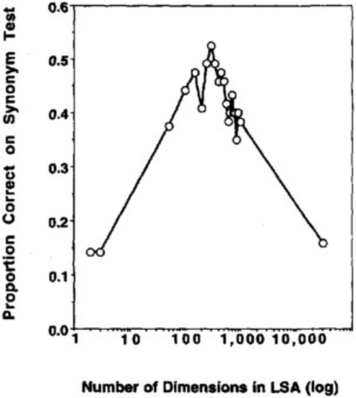
The answer choice with the smallest angle wins.

The 1997 Results

In short, they found that with the proper dimension optimization, they can score about **half the questions correct** (after controlling for chance guessing $\frac{\text{corr}-\text{chance}}{1-\text{chance}}$) which is good enough to get into college!

Also, there is a local maximum with too little dimensions not accurately representing the space and too many makes for a space too diffuse for interrelationships to be exploited.

Other conclusions as well...



My Study — A Duplication with a Twist

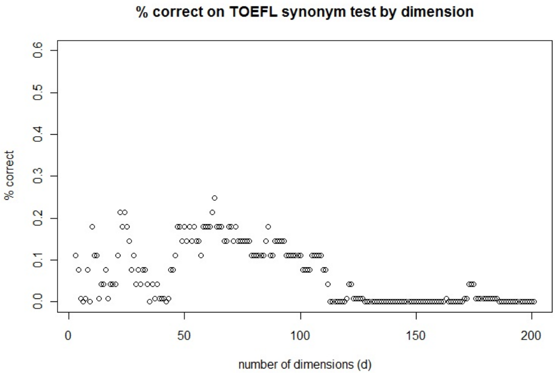
I have attempted to duplicate their analysis using Internet 4-grams from Google's 2006 corpus.

Due to computational limitations, I used the following parameters:

$$\begin{aligned} p &= 2000 \quad \Rightarrow \quad n = 2173 \\ d_{MAX} &= 200 \\ \text{MAX\_TOEFL} &= 5 \end{aligned}$$

Only 39 of the 80 TOEFL questions were able to be tested, and each not completely. Hence my data is very preliminary.

My Results



We can see optimal dimensionality between 50-60 dimensions and a precipitous drop after 110.

Conclusions and Future Idea

Even with this exceedingly raw setup, we can be convinced of LSA working using merely 4-grams. That means there's a lot of information in only four words if enough of them are read.

Future Idea

Optimize over both length of context and dimensionality of context — where is cusp? At what point do we get diminiship returns

Acknowledgments

Dean Foster
Thomas Landauer Lab (for providing TOEFL data)

Funding from GRFP

Thanks for letting me present today!