How Document Space is like an Elephant

David J. Marchette
dmarchette@gmail.com

Naval Surface Warfare Center
Code B10

25th January 2006

WARFARE CENTERS
DAHLGREN

Document Space

What is **Document Space**? My talk will discuss this. I won't:

- Tell you about my particular algorithm.
- Talk about a specific sub-problem in text mining (although there will be a bias toward the types of problems I am interested in).
- Argue for or against any particular methodology.

I take a pattern recognition/data mining perspective. I will try to:

- Take a small step towards defining "Document Space".
- Entertain you while I do it.

Other Possible Answers

Document space might be a space of:

- probabilistic models (such as we saw yesterday morning).
- language models.
- syntactic/parsing rules or methods.

I will be considering it more as a mapping

$$F : \text{documents} \rightarrow \mathbb{R}^d.$$

In fact, I'll try to argue that it is a space of mappings.

Outline

Pattern Recognition

Blind Men and Elephants

Corpus Dependence

Conditioning

More Elephants

Outline

Pattern Recognition

Blind Men and Elephants

Corpus Dependence

Conditioning

More Elephants

Outline

Pattern Recognition

Blind Men and Elephants

Corpus Dependence

Conditioning

More Elephants

Outline

- Pattern Recognition
- Blind Men and Elephants
- Corpus Dependence
- Conditioning
- More Elephants

NAVSEA WARFARE CENTERS DAHLGREN

Outline

Pattern Recognition

Blind Men and Elephants

Corpus Dependence

Conditioning

More Elephants

Outline

Pattern Recognition

Blind Men and Elephants

Corpus Dependence

Conditioning

More Elephants

WARFARE CENTERS
DAHLGREN

The Steps of Pattern Recognition

1. Extract features from a signal.
   - Usually done somewhat ad hoc.
   - Requires an expert in the application to do it right.
   - Depends on what the task is.
2. Select and project the features into some space ($\mathbb{R}^d$).
3. Build the exploitation algorithm (classification, regression, clustering). (Choose a model.)
4. Evaluate the algorithm.
5. Publish, or start over at one of the above steps.

Document Space Lives Here

1. Extract features from a signal.
2. Select and project the features into some space ($\mathbb{R}^d$).

- What is "Document Space"?
- It is a mapping from documents (a corpus) into some numerical quantities

$$F : \mathcal{C} \longrightarrow X$$

- then a mapping from the quantities into a space in which to perform pattern recognition

$$P : X \longrightarrow \mathbb{R}^d$$

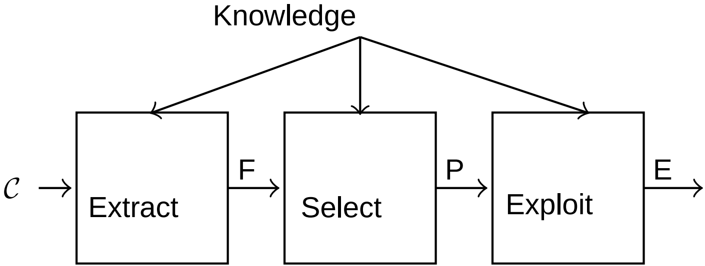
This need not really be $\mathbb{R}^d$, but almost always is in practice.

Pattern Recognition

$$\mathcal{C} \rightarrow \boxed{\text{Extract}} \xrightarrow{\text{F}} \boxed{\text{Select}} \xrightarrow{\text{P}} \boxed{\text{Exploit}} \xrightarrow{\text{E}}$$

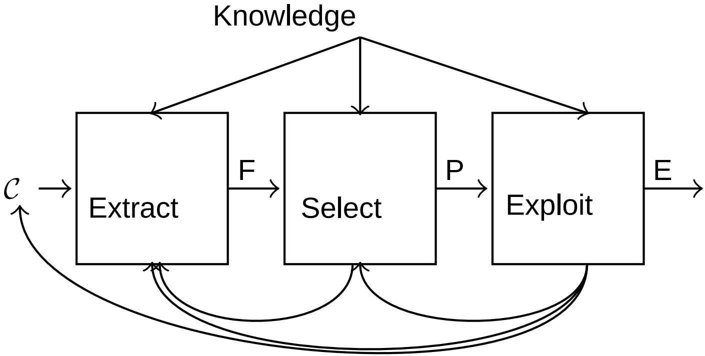
Extract features, select/project, and exploit.

Pattern Recognition



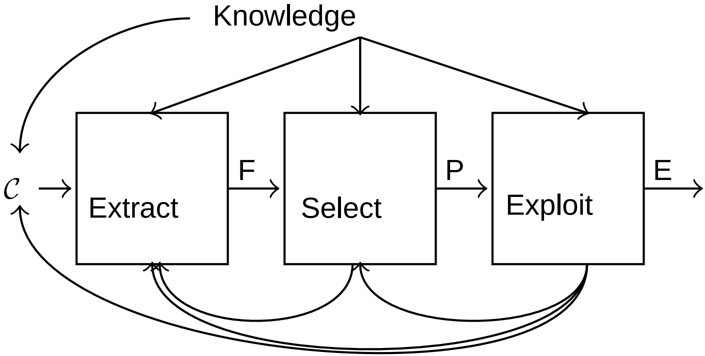
These all require input of knowledge at each stage.

Pattern Recognition



Data analysis feeds back to the modules.

Pattern Recognition



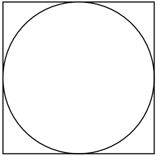
The user may also choose to focus on a subset of the corpus.

Lessons from Pattern Recognition

- Do not extract more than you need to perform the task. (The **Curse of Dimensionality**)
  - This advice is usually ignored, and redundancy and noise is removed in the dimensionality reduction stage.
- Consult an expert. Also often ignored.
- Operate in as low a dimensional space as you can (but no lower).
- Use the simplest model (but no simpler).
- Iteration wins the race.

The Curse of Dimensionality

Consider a circle inscribed within a square:



- Most of the volume of the square is in the circle. This is what we think the world is like.
- Now consider a (high dimensional) sphere inscribed in a hypercube. As the dimensionality increases, the ratio of the volume of the sphere to that of the cube goes to zero!
- All the “stuff” is in the corners (along the edges)!

NAVSEA WARFARE CENTERS DAHLGREN

Curse of Dimensionality Revisited

Suppose the data (in $\mathbb{R}^d$) are distributed normally (iid), with (known) identity covariance, and means:

$$\pm\left(1, \frac{1}{\sqrt{2}}, \ldots, \frac{1}{\sqrt{d}}\right).$$

We classify a new observation by computing the distance to these means, and assigning the class according to the closest.

- If we know the means a priori, the probability of error goes to zero as the dimension increases.
- If we do not, and have to estimate them, the probability of error goes to 0.5: chance.
- Adding features (even ones with discriminant information) does not necessarily make the classifier better.

What Is Document Space?

A mapping

$$F(\cdot|\mathcal{P},\mathcal{K}) : \mathcal{C} \to \mathbb{R}^d$$

Rather than asking

- shouldn't that be something other than Euclidean space? (Carey)
  or
- what should d be? (Carey)

I propose that the key is understanding the relationship between F and $\mathcal{P}$ (the "problem") and $\mathcal{K}$ ("knowledge").

NAVSEA WARFARE CENTERS DAHLGREN

Outline

Pattern Recognition

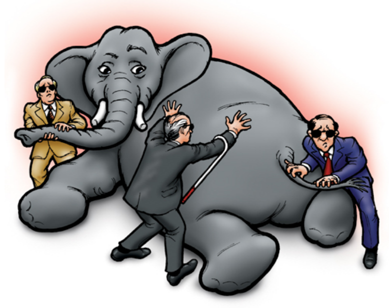
Blind Men and Elephants

Corpus Dependence

Conditioning

More Elephants

You Look For What You Know



www.silentera.com/CBD/elephant.html

- Information theorests see word-count histograms.
- Speech recognition people see Hidden Markov Models.
- Natural language processors see nouns and verbs and syntax and ontologies.

Knowledge Directly Impacts Feature Extraction

- Stemming, stop word removal, noise reduction.
- Word counts, mutual information, ngrams, TFIDF.
- Tagging and sentence structure.
- Semantics, thesauri.

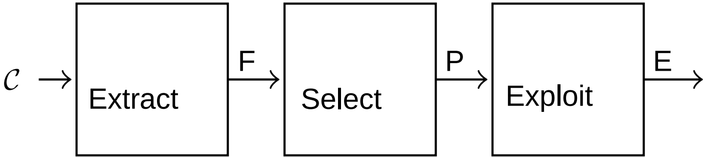
The Problem Directly Impacts Exploitation

- Classification vs Clustering
- Understanding vs Comparison
- Translation
- Summarization

What is done with the features depends on the problem.

NAVSEA WARFARE CENTERS DAHLGREN

Document Space (First Cut)



$$\mathcal{F}(\cdot|\mathcal{P},\mathcal{K}) = \mathcal{F}(\cdot, F_{\mathcal{K}}, P_{\mathcal{P},\mathcal{K}}, E_{\mathcal{P}})$$

Maybe document space is feature extraction from documents into some feature space. What properties should this have?

Outline

- Pattern Recognition
- Blind Men and Elephants
- Corpus Dependence
- Conditioning
- More Elephants

Corpus Dependence

- The words/phrases/constructs that are important depend on the corpus.
- Change the corpus and you change “document space”.
- If you think of “document space” as the space into which documents project, it is not independent of the corpus.

NAVSEA WARFARE CENTERS DAHLGREN

Importance is Relative

What are the important words/phrases in the following?

A Gebusi woman in New Guinea, decked out in her dance costume, sleeps on a woodpile during a male initiation ceremony.

This is a sentence from a document in a corpus I have worked with.

Importance is Relative

What are the important words/phrases in the following?

A **Gebusi** woman in **New Guinea**, decked out in her **dance costume**, sleeps on a woodpile during a **male initiation ceremony**.

In the absence of any a priori knowledge (beyond our knowledge of English), we might choose these.

Importance is Relative

What are the important words/phrases in the following?

A Gebusi woman in New Guinea, decked out in her **dance costume**, sleeps on a woodpile during a male initiation **ceremony**.

If we knew the corpus was about fashion, we might choose these.

Importance is Relative

What are the important words/phrases in the following?

A Gebusi woman in New Guinea, decked out in her dance costume, **sleeps** on a woodpile during a male initiation ceremony.

If we knew the corpus was about the study of sleep, this might be important.

Importance is Relative

The important words depend on the context of the rest of the document, and the rest of the corpus. This was from an article on sleep, in a corpus of science articles, including anthropology and medicine.

A **Gebusi woman** in **New Guinea**, decked out in her dance costume, sleeps on a woodpile during a male initiation ceremony.

NAVSEA WARFARE CENTERS DAHLGREN

Corpus Dependence

- Mapping documents into “Document Space” requires extracting information from the documents within the corpus.
- What words/phrases correspond to **information** depends on the corpus.
- Property one for document space: corpus dependent feature extraction.

Outline

Conditioning

There are two main types of conditioning of interest:

1. Conditioning on the corpus.
2. Conditioning on the problem.

- In the first, we have corpus-dependent features: important words (feature extraction, projections) depend on the context of the overall corpus.
- In the second, we incorporate a priori knowledge of what is important.

Corpus Dependent Feature Extraction

Features depend on:

- The words we use — thresholding to remove “unimportant” words.
- The topics in the corpus — the specific documents or document classes.

I will illustrate these on a small data set from Science News.

Science News Data

- 1160 articles from Science News — typically one page long.
- Classified into 8 categories:
  Anthropology, Astronomy, Behavior, Earth Sciences, Life Sciences, Math and Computer Science, Medicine, Physical Sciences.
- Categorization performed by a single individual reading the documents (clearly many documents can have several classes).
- A random subset of 50 from each category chosen.

Mutual Information

- We need to map the documents into some space for processing.
- For illustration purposes, I'll map into $\mathbb{R}^2$ and use scatterplots for visualization.
- We need to extract information from the documents:
  - Word count histogram.
  - Mutual information.
- Then a method for comparing documents: Cosine-dissimilarity.
- Finally embed in $\mathbb{R}^2$: multidimensional scaling.

Mutual Information

Let $c_{w,d}$ be the number of times that the word w has occurred in the document d and let $N_{\mathcal{C}}$ be the total number of words (counting duplicates) in the corpus $\mathcal{C}$. Let $f_{w,d} = c_{w,d}/N_{\mathcal{C}}$. Then the mutual information between document d and word w is given by

$$m_{w,d}^{\mathcal{C}} = \log\left(\frac{f_{w,d}}{\sum_{z\in\mathcal{C}} f_{w,z} \sum_i f_{i,d}}\right)$$

Let N_d be the number of words (counting duplicates) in document d. Let $c_{w,\mathcal{C}}$ be the number of times that the word w appears in the corpus $\mathcal{C}$.

$$m_{w,d}^{\mathcal{C}} = \log\left(\frac{\frac{c_{w,d}}{N_d}}{\frac{c_{w,\mathcal{C}}}{N_{\mathcal{C}}}}\right)$$

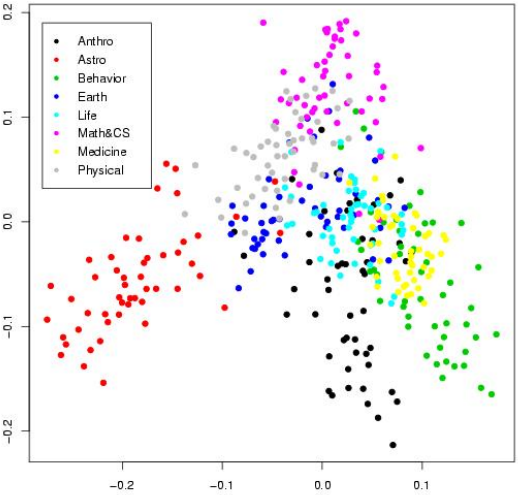
Interpoint Distances and Projections

Dissimilarities can be computed via cosine distance: Let a and b be documents, represented by a vector of mutual informations corresponding to each word in the lexicon.
The cosine-dissimilarity:

$$\rho(a,b) = 1 - \frac{a \cdot b}{|a||b|}$$

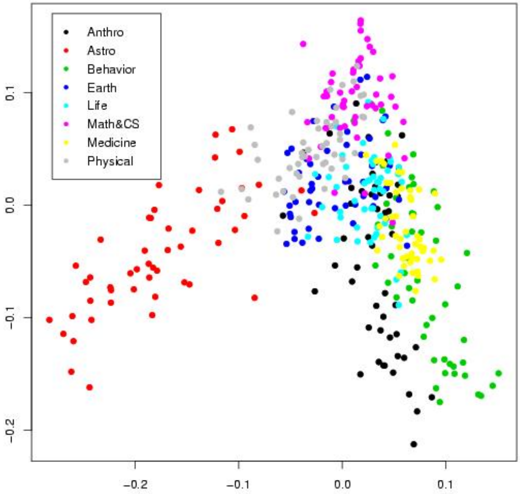
We set ρ to be a large number (≥ 2) if the documents share no words.
Note that $\rho(a,b) = 1 - \frac{|S_a \cap S_b|}{\sqrt{|S_a||S_b|}}$ if we use a binary threshold.
We can project the data to $\mathbb{R}^2$ for visualization via multidimensional scaling.

Science News



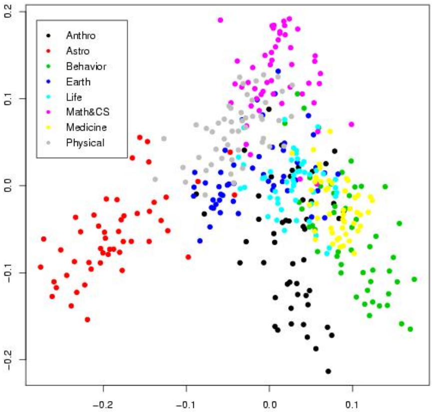
Threshold: 0.

Removing “Unimportant” Words

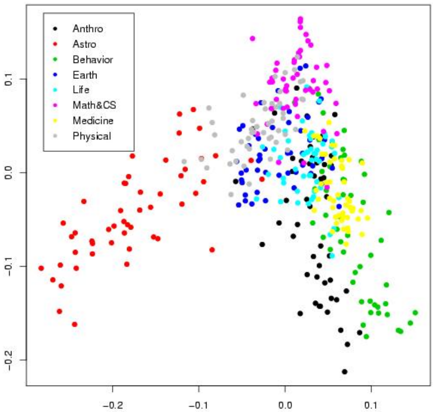


Threshold: 2.

Side-By-Side Comparison: Unimportant Words



Threshold 0



Threshold 2

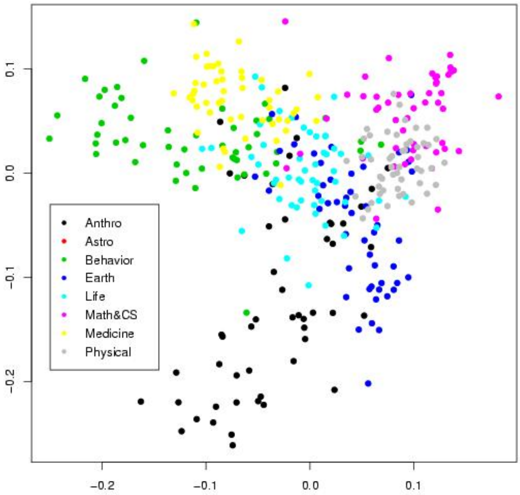
WARFARE CENTERS
DAHLGREN

Stop Words

- Stop words can be removed (*a, the, and, by* . . .).
- Some words are always content-free, but “unimportant” words are context (corpus) dependent.
- This is feature selection/dimensionality reduction.

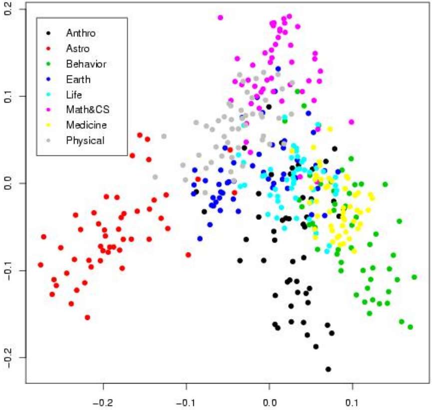
Feature selection is corpus dependent.

Context: Removing Astronomy

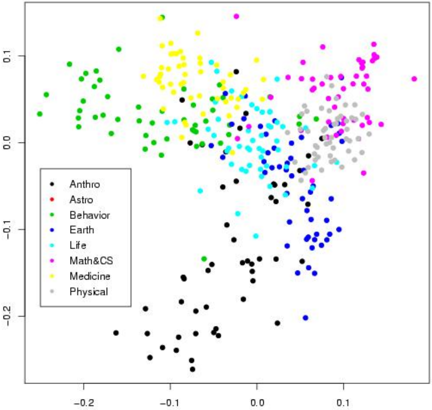


Threshold: 0.

Side-By-Side Comparison: Astronomy



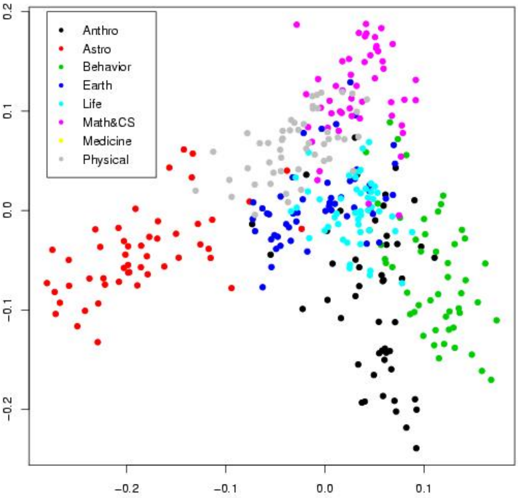
Threshold 0



Threshold 0

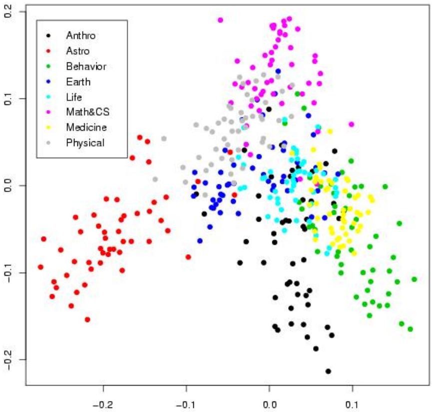
NAVSEA WARFARE CENTERS DAHLGREN

Context: Removing Medicine

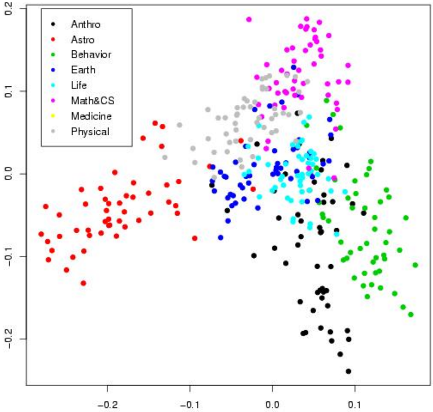


Threshold: 0.

Side-By-Side Comparison: Medicine



Threshold 0



Threshold 0

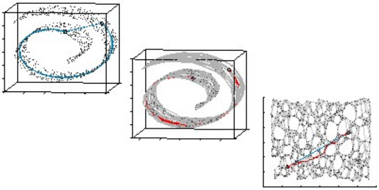
Corpus Dependence

Feature selection depends on the corpus.

- The documents (topics, classes) determine the “important words”.
- These drive the feature selection/projection.
- A document will “look different” within the context of one corpus compared to that of another: with/without astronomy.

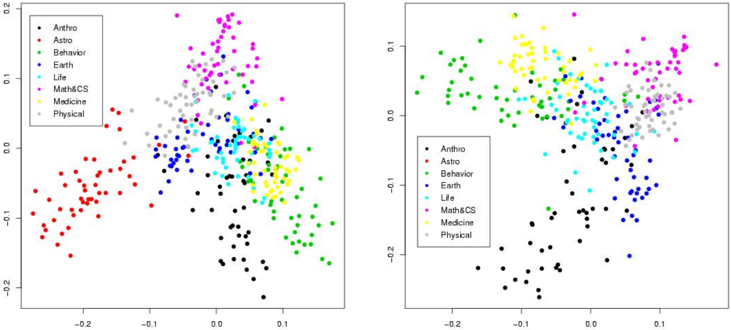
Why Isn't This Local Dimensionality Reduction?

Standard picture from manifold learning (Isomap, LLE, etc):



Note: Different projection (features) in different positions in feature space.

The Space Depends on the Corpus



Far away changes effect local projections.

Face Recognition Example

Different features may be extracted for different situations.

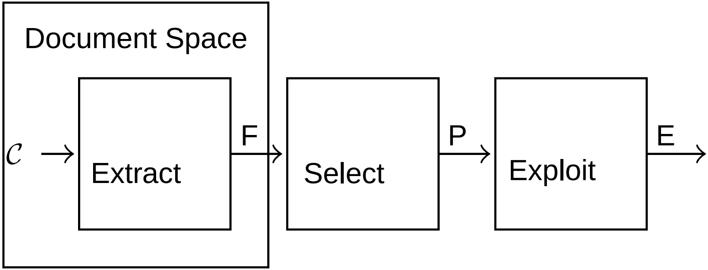
What Is the Problem?

- Recognize the person.
- Detect the face.
- Detect words.
- Distinquish between student and professor.
- Determine inside vs outside.
- Determine light source.

Different features may be extracted for different problems.

Property two: document space depends on the problem (many document spaces?).

Each Step is Conditional



$$
\begin{aligned}
\mathcal{F}(\cdot|\mathcal{P},\mathcal{K}) &= \mathcal{F}(\cdot, F(\cdot|\mathcal{K},\mathcal{P},\mathcal{C}), P(\cdot|\mathcal{P}), E(\cdot|\mathcal{P})) \\
&= \text{Document Space},\ \text{Work Space},\ \text{Algorithm}
\end{aligned}
$$

WARFARE CENTERS
DAHLGREN

Is $\mathbb{R}^d$ the Right Model?

- Psychologists would say no:
  - Plenty of examples of $A > B$ and $B > C$ and $C > A$.
- Psychologists would say yes:
  - Major users of PCA/MDS.
- A case can be made that the initial space should not be Euclidean.
- I think this is less of an issue than that of determining how to characterize corpus/problem dependent feature extraction.

Outline

Pattern Recognition

Blind Men and Elephants

Corpus Dependence

Conditioning

More Elephants

More Elephants

Feedback refines the feature extraction:

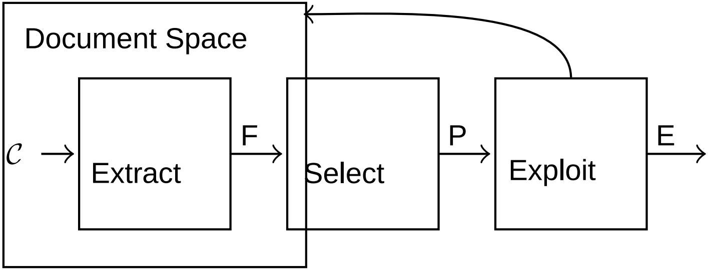
Once we determine we have an elephant, the question becomes: Which species?

Text Processing

- Once we:
  - Determine the important words
  - Determine the sub-corpus of interest
  - Refine our knowledge of the problem

  we revisit the feature extraction problem.
- We may start with bag-of-words, use this to determine important words, a subcorpus of interest, etc, then do syntactic or semantic analysis on subsets of the words/documents.
- Document space (which features we extract) depends on previous iterations of the process.
- Document space depends on the problem and feedback modifies the problem, as well as the features.

Feedback



$$
\begin{aligned}
\mathcal{F}(\cdot|\mathcal{P},\mathcal{K}) &= \mathcal{F}(\cdot, F(\cdot|\mathcal{K},\mathcal{C},\mathcal{P}), P(\cdot|F,\mathcal{P}), E(\cdot|P)) \\
&= \text{Document Space},\ \text{Work Space},\ \text{Algorithm}
\end{aligned}
$$

Some of feature selection/projection is corpus/problem dependent.
Feedback needs to be incorporated in the model.

NAVSEA WARFARE CENTERS DAHLGREN

Statistical Inference

- Some people like to frame everything as a hypotheses test:

$$
\begin{aligned}
H_0 &: \quad \text{no signal} \\
H_1 &: \quad \text{some signal}
\end{aligned}
$$

- The problem is, we don't know what the alternative should be.
- A general alternative produces a test with no power.
- If we could specify the "correct" alternative, we could design a more powerful test.

Why Not Just Answer the Question?

- Imagine typing “What is the distance between San Diego and New York?” into a search engine.
- You want it to return a document containing the answer.
- IPAM typed in the question “What is Document Space” and got back
  - Generative models.
  - Language models.
  - Pattern recognition models.
  - Problem specific discussion.
  - Data analysis methodologies applied to text.
- These are all related to the question, and point out that the question is too broad.
- Also, the answer is not yet known.

NAVSEA WARFARE CENTERS DAHLGREN

What is Document Space?

- Document space is a collection of maps taking documents into some numerical (dissimilarity?) space:

$$F(\cdot|\mathcal{P},\mathcal{K},\mathcal{C}) : D \to \mathbb{R}^d$$

- Making this precise requires some handle on
  - What type of mathematical objects are $\mathcal{P}$ and $\mathcal{K}$.
  - How is feedback (refining the problem, reducing the corpus, changing the features—reselecting the function F) to be thought about?

WARFARE CENTERS
DAHLGREN

What is Document Space?

Document space is the space of embeddings, not the embedding in space. It has the properties of:

1. Corpus dependent features.
2. Dependence on the problem (or different spaces for different problems).
3. A refinenment (feedback).
4. Undoubtedly lots of others.

What is Document Space?

Document space is the space of embeddings, not the embedding in space. It has the properties of:

1. Corpus dependent features.
2. Dependence on the problem (or different spaces for different problems).
3. A refinenment (feedback).
4. Undoubtedly lots of others.

Not unlike elephants.