DAT630
Retrieval Models II.

Search Engines, Chapters 7

11/10/2016

Krisztian Balog | University of Stavanger

General Scoring Formula

$$score(d, q) = \sum_{t \in q} w_{t,d} \cdot w_{t,q}$$

Relevance score
It is computed for each document *d* in the collection for a given input query *q*

Documents are returned in decreasing order of this score

It is enough to consider terms in the query

Term's weight in the document

Term's weight in the query

Language Models

Language Models

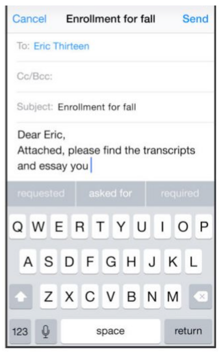
- Based on the notion of probabilities and processes for generating text

Uses

- Speech recognition
  - "I ate a cherry" is a more likely sentence than "Eye eight uh Jerry"
- OCR & Handwriting recognition
  - More probable sentences are more likely correct readings
- Machine translation
  - More likely sentences are probably better translations

Uses

- Completion prediction
  - Please turn off your cell _____
  - Your program does not ______
- *Predictive text input systems* can guess what you are typing and give choices on how to complete it



Ranking Documents using Language Models

- Represent each document as a multinomial probability distribution over terms
- Estimate the probability that the query was "generated" by the given document
  - "How likely is the search query given the language model of the document?"

Standard Language Modeling approach

- Rank documents d according to their likelihood of being relevant given a query q: $P(d|q)$

$$P(d|q) = \frac{P(q|d)P(d)}{P(q)} \propto P(q|d)P(d)$$

Query likelihood
Probability that query q was "produced" by document d

Document prior
Probability of the document being relevant to *any* query

$$P(q|d) = \prod_{t \in q} P(t|\theta_d)^{f_{t,q}}$$

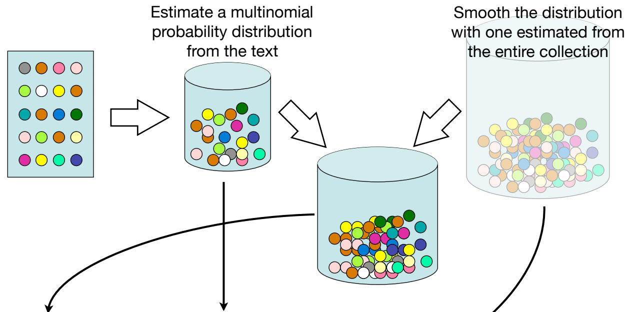
Standard Language Modeling approach (2)

$$P(q|d) = \prod_{t \in q} \underbrace{P(t|\theta_d)}_{\substack{\textbf{Document language model}\\ \text{Multinomial probability distribution}\\ \text{over the vocabulary of terms}}}^{\overbrace{f_{t,q}}^{\text{Number of times } t \text{ appears in } q}}$$

$$P(t|\theta_d) = (1-\lambda)\underbrace{P(t|d)}_{\substack{\textbf{Empirical document model}\\ \frac{f_{t,d}}{|d|}}} + \overbrace{\lambda}^{\textbf{Smoothing parameter}}\underbrace{P(t|C)}_{\substack{\textbf{Collection model}\\ \frac{\sum_{d'} f_{t,d'}}{\sum_{d'} |d'|}}}$$

Maximum likelihood estimates

Language Modeling



$$\overbrace{P(t|\theta_d)} = (1-\lambda)\overbrace{P(t|d)} + \lambda\underbrace{P(t|C)}$$

Example

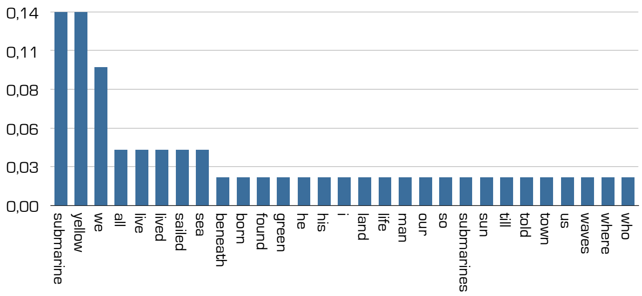
In the town where I was born,
Lived a man who sailed to sea,
And he told us of his life,
In the land of submarines,

So we sailed on to the sun,
Till we found the sea green,
And we lived beneath the waves, In our yellow submarine,

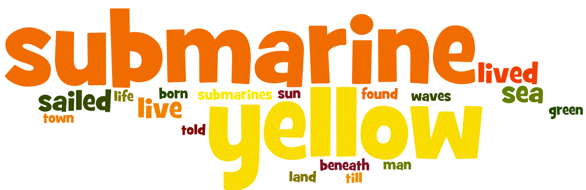
We all live in yellow submarine, yellow submarine, yellow submarine, We all live in yellow submarine, yellow submarine, yellow submarine.

Empirical document LM

$$P(t|d) = \frac{f_{t,d}}{|d|}$$



Alternatively...



Scoring a query

$$q = \{\text{sea}, \text{submarine}\}$$

$$P(q|d) = P(\text{“sea”}|\theta_d) \cdot P(\text{“submarine”}|\theta_d)$$

Scoring a query

$$q = \{\text{sea}, \text{submarine}\}$$

$$P(q|d) = \overset{0.03602}{P(\text{“sea”}|\theta_d)} \cdot P(\text{“submarine”}|\theta_d)$$

$$\overset{0.9}{(1-\lambda)}\overset{0.04}{P(\text{“sea”}|d)} + \overset{0.1}{\lambda}\overset{0.0002}{P(\text{“sea”}|C)}$$

t	P(t|d)
submarine	0,14
sea	0,04
...	

t	P(t|C)
submarine	0,0001
sea	0,0002
...	

Scoring a query

$$q = \{\text{sea}, \text{submarine}\}$$

$$\overset{0.04538}{P(q|d)} = \overset{0.03602}{P(\text{“sea”}|\theta_d)} \cdot \overset{0.12601}{P(\text{“submarine”}|\theta_d)}$$

$$\overset{0.9}{(1-\lambda)}\overset{0.14}{P(\text{“submarine”}|d)} + \overset{0.1}{\lambda}\overset{0.0001}{P(\text{“submarine”}|C)}$$

t	P(t|d)
submarine	0,14
sea	0,04
...	

t	P(t|C)
submarine	0,0001
sea	0,0002
...	

Smoothing

- **Jelinek-Mercer smoothing**

$$P(t|\theta_d) = (1-\lambda)P(t|d) + \lambda P(t)$$

  - Smoothing parameter is λ
  - Same amount of smoothing is applied to all documents

- **Dirichlet smoothing**

$$p(t|\theta_d) = \frac{f_{t,d} + \mu \cdot p(t)}{|d| + \mu}$$

  - Smoothing parameter is μ
  - Smoothing is inversely proportional to the document length

Relation between Smoothing Methods

- Jelinek Mercer:

$$P(t|\theta_d) = (1-\lambda)P(t|d) + \lambda P(t)$$

- by setting:

$$(1-\lambda) = \frac{|d|}{|d| + \mu} \qquad \lambda = \frac{\mu}{|d| + \mu}$$

- Dirichlet:

$$p(t|\theta_d) = \frac{f_{t,d} + \mu \cdot p(t)}{|d| + \mu}$$

Practical Considerations

- Since we are multiplying small probabilities, it's better to perform computations in the log space

$$P(q|d) = \prod_{t \in q} P(t|\theta_d)^{f_{t,q}}$$

$$\underbrace{\log P(q|d)}_{score(d,q)} = \sum_{t \in q} \underbrace{\log P(t|\theta_d)}_{w_{t,d}} \cdot \underbrace{f_{t,q}}_{w_{t,q}}$$

$$score(d,q) = \sum_{t \in q} w_{t,d} \cdot w_{t,q}$$

Exercise

Exercise

GitHub: exercises/20161011-sol.xlsx

	term frequencies					empirical language models					collection language model	smoothed language models				
term	D1	D2	D3	D4	D5	D1	D2	D3	D4	D5		D1	D2	D3	D4	D5
T1		1			1	0	0,2	0	0	0,25	0,091	0,009	0,189	0,009	0,009	0,234
T2		1			1	0	0,2	0	0	0,25	0,091	0,009	0,189	0,009	0,009	0,234
T3	3	2	2		1	0,6	0,4	0,5	0	0,25	0,364	0,576	0,396	0,486	0,036	0,261
T4			1	1		0	0	0,25	0,25	0	0,091	0,009	0,009	0,234	0,234	0,009
T5			1	1	1	0	0	0,25	0,25	0,25	0,136	0,014	0,014	0,239	0,239	0,239
T6	2	1		2		0,4	0,2	0	0,5	0	0,227	0,383	0,203	0,023	0,473	0,023
|D|	5	5	4	4	4	1	1	1	1	1	1	1	1	1	1	1
Jelinek-Mercer smoothing																
smoothing parameter						0,1										

Exercise

	term frequencies					empirical language models					collection language model	smoothed language models				
term	D1	D2	D3	D4	D5	D1	D2	D3	D4	D5		D1	D2	D3	D4	D5
T1		1			1	0	0,2	0	0	0,25	0,091	0,009	0,189	0,009	0,009	0,234
T2		1			1	0	0,2	0	0	0,25	0,091	0,009	0,189	0,009	0,009	0,234
T3	3	2	2		1	0,6	0,4	0,5	0	0,25	0,364	0,576	0,396	0,486	0,036	0,261
T4			1	1		0	0	0,25	0,25	0	0,091	0,009	0,009	0,234	0,234	0,009
T5			1	1	1	0	0	0,25	0,25	0,25	0,136	0,014	0,014	0,239	0,239	0,239
T6	2	1		2		0,4	0,2	0	0,5	0	0,227	0,383	0,203	0,023	0,473	0,023
|D|	5	5	4	4	4	1	1	1	1	1	1	1	1	1	1	1
Jelinek-Mercer smoothing																
smoothing parameter						0,1										

$$P(t|\theta_d) = (1-\lambda)\underbrace{P(t|d)}_{} + \lambda P(t|C)$$

Document language model computation

Exercise

	term frequencies					empirical language models					collection language model	smoothed language models				
term	D1	D2	D3	D4	D5	D1	D2	D3	D4	D5		D1	D2	D3	D4	D5
T1		1			1	0	0,2	0	0	0,25	0,091	0,009	0,189	0,009	0,009	0,234
T2		1			1	0	0,2	0	0	0,25	0,091	0,009	0,189	0,009	0,009	0,234
T3	3	2	2		1	0,6	0,4	0,5	0	0,25	0,364	0,576	0,396	0,486	0,036	0,261
T4			1	1		0	0	0,25	0,25	0	0,091	0,009	0,009	0,234	0,234	0,009
T5			1	1	1	0	0	0,25	0,25	0,25	0,136	0,014	0,014	0,239	0,239	0,239
T6	2	1		2		0,4	0,2	0	0,5	0	0,227	0,383	0,203	0,023	0,473	0,023
|D|	5	5	4	4	4	1	1	1	1	1	1	1	1	1	1	1
Jelinek-Mercer smoothing																
smoothing parameter						0,1										

$$P(t|\theta_d) = (1-\lambda)P(t|d) + \lambda \underbrace{P(t|C)}_{}$$

Document language model computation

Exercise

	term frequencies					empirical language models					collection language model	smoothed language models				
term	D1	D2	D3	D4	D5	D1	D2	D3	D4	D5		D1	D2	D3	D4	D5
T1		1			1	0	0,2	0	0	0,25	0,091	0,009	0,189	0,009	0,009	0,234
T2		1			1	0	0,2	0	0	0,25	0,091	0,009	0,189	0,009	0,009	0,234
T3	3	2	2		1	0,6	0,4	0,5	0	0,25	0,364	0,576	0,396	0,486	0,036	0,261
T4			1	1		0	0	0,25	0,25	0	0,091	0,009	0,009	0,234	0,234	0,009
T5			1	1	1	0	0	0,25	0,25	0,25	0,136	0,014	0,014	0,239	0,239	0,239
T6	2	1		2		0,4	0,2	0	0,5	0	0,227	0,383	0,203	0,023	0,473	0,023
|D|	5	5	4	4	4	1	1	1	1	1	1	1	1	1	1	1
Jelinek-Mercer smoothing																
smoothing parameter						0,1										

$$P(t|\theta_d) = (1-\underbrace{\lambda})P(t|d) + \lambda P(t|C)$$

Document language model computation

Exercise

	term frequencies					empirical language models					collection language model	smoothed language models				
term	D1	D2	D3	D4	D5	D1	D2	D3	D4	D5		D1	D2	D3	D4	D5
T1		1			1	0	0,2	0	0	0,25	0,091	0,009	0,189	0,009	0,009	0,234
T2		1			1	0	0,2	0	0	0,25	0,091	0,009	0,189	0,009	0,009	0,234
T3	3	2	2		1	0,6	0,4	0,5	0	0,25	0,364	0,576	0,396	0,486	0,036	0,261
T4			1	1		0	0	0,25	0,25	0	0,091	0,009	0,009	0,234	0,234	0,009
T5			1	1	1	0	0	0,25	0,25	0,25	0,136	0,014	0,014	0,239	0,239	0,239
T6	2	1		2		0,4	0,2	0	0,5	0	0,227	0,383	0,203	0,023	0,473	0,023
|D|	5	5	4	4	4	1	1	1	1	1	1	1	1	1	1	1
Jelinek-Mercer smoothing																
smoothing parameter						0,1										

$$\underbrace{P(t|\theta_d)} = (1-\lambda)P(t|d) + \lambda P(t|C)$$

Document language model computation

Exercise

	term frequencies					empirical language models					collection language model	smoothed language models				
term	D1	D2	D3	D4	D5	D1	D2	D3	D4	D5		D1	D2	D3	D4	D5
T1		1			1	0	0,2	0	0	0,25	0,091	0,009	0,189	0,009	0,009	0,234
T2		1			1	0	0,2	0	0	0,25	0,091	0,009	0,189	0,009	0,009	0,234
T3	3	2	2		1	0,6	0,4	0,5	0	0,25	0,364	0,576	0,396	0,486	0,036	0,261
T4			1	1		0	0	0,25	0,25	0	0,091	0,009	0,009	0,234	0,234	0,009
T5			1	1	1	0	0	0,25	0,25	0,25	0,136	0,014	0,014	0,239	0,239	0,239
T6	2	1		2		0,4	0,2	0	0,5	0	0,227	0,383	0,203	0,023	0,473	0,023
|D|	5	5	4	4	4	1	1	1	1	1	1	1	1	1	1	1
Jelinek-Mercer smoothing																
smoothing parameter						0,1										
q="T3"												0,576	0,396	0,486	0,036	0,261
q="T2 T1"												0,000	0,036	0,000	0,000	0,055
q="T6"												0,383	0,203	0,023	0,473	0,023
q="T3 T1 T3 T2"												0,000	0,006	0,000	0,000	0,004

Scoring a query

$$P(q|d) = \prod_{t \in q} P(t|\theta_d)^{f_{t,q}}$$

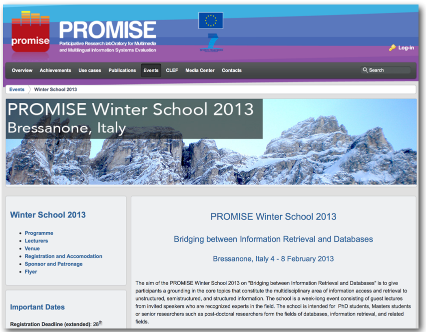
$\underbrace{\text{P(q="T2 T1"|D2)}} = \underbrace{\text{P(T2|D2)}} * \underbrace{\text{P(T1|D2)}}$

Fielded Variants

Motivation

- Documents are composed of multiple fields
  - E.g., title, body, anchors, etc.
- Modeling internal document structure may be beneficial for retrieval

Example



Unstructured representation

```
PROMISE Winter School 2013
Bridging between Information Retrieval and Databases
Bressanone, Italy 4 - 8 February 2013
The aim of the PROMISE Winter School 2013 on "Bridging between Information
Retrieval and Databases" is to give participants a grounding in the core
topics that constitute the multidisciplinary area of information access and
retrieval to unstructured, semistructured, and structured information. The
school is a week-long event consisting of guest lectures from invited
speakers who are recognized experts in the field. The school is intended for
PhD students, Masters students or senior researchers such as post-doctoral
researchers form the fields of databases, information retrieval, and related
fields.
[...]
```

```
<html>
<head>
  <title>Winter School 2013</title>
  <meta name="keywords" content="PROMISE, school, PhD, IR, DB, [...]" />
  <meta name="description" content="PROMISE Winter School 2013, [...]" />
</head>
<body>
  <h1>PROMISE Winter School 2013</h1>
  <h2>Bridging between Information Retrieval and Databases</h2>
  <h3>Bressanone, Italy 4 - 8 February 2013</h3>
  <p>The aim of the PROMISE Winter School 2013 on "Bridging between
  Information Retrieval and Databases" is to give participants a grounding
  in the core topics that constitute the multidisciplinary area of
  information access and retrieval to unstructured, semistructured, and
  structured information. The school is a week-long event consisting of
  guest lectures from invited speakers who are recognized experts in the
  field. The school is intended for  PhD students, Masters students or
  senior researchers such as post-doctoral researchers form the fields of
  databases, information retrieval, and related fields. </p>
  [...]
</body>
</html>
```

Fielded representation

based on HTML markup

```
title:    Winter School 2013

meta:     PROMISE, school, PhD, IR, DB, [...]
          PROMISE Winter School 2013, [...]

headings: PROMISE Winter School 2013
          Bridging between Information Retrieval and Databases
          Bressanone, Italy 4 - 8 February 2013

body:     The aim of the PROMISE Winter School 2013 on "Bridging between
          Information Retrieval and Databases" is to give participants a
          grounding in the core topics that constitute the multidisciplinary
          area of information access and retrieval to unstructured,
          semistructured, and structured information. The school is a week-
          long event consisting of guest lectures from invited speakers who
          are recognized experts in the field. The school is intended for
          PhD students, Masters students or senior researchers such as post-
          doctoral researchers form the fields of databases, information
          retrieval, and related fields.
```

In Web Search: Links

- Links are a key component of the Web
- Important for navigation, but also for search
  - Both the anchor text and the destination link are used by search engines

```
<a href="http://example.com">Example website</a>
```

Destination link: `http://example.com` — Anchor text: `Example website`

Anchor Text

- Anchor text tends to be short, descriptive, and similar to query text
- Usually written by people who are not the authors of the destination page
  - Can describe a destination page from a different perspective, or emphasize the most important aspect of the page from a community viewpoint

Anchor Text

- Collection of anchor text in all links pointing to a given page are used as a description of the content of the destination page
  - I.e., added as an additional document field
- Retrieval experiments have shown that anchor text has significant impact on effectiveness for *some types of queries*
  - Essential for searches where the user is trying to find a homepage for a particular topic, person, or organization

Anchor Text

page1

```
I'll be presenting our work at a
<a href="pageX">winter school</a>
in Bressanone, Italy.
```

page2

```
List of winter schools in 2013:
<ul>
<li><a href="pageX">information
retrieval</a></li>
…
</ul>
```

page3

```
The PROMISE Winter School in will
feature a range of <a
href="pageX">IR lectures</a> by
experts from the field.
```

Anchor Text

page1

```
I'll be presenting our work at a
<a href="pageX">winter school</a>
in Bressanone, Italy.
```

page2

```
List of winter schools in 2013:
<ul>
<li><a href="pageX">information
retrieval</a></li>
…
</ul>
```

page3

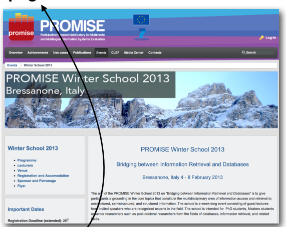
```
The PROMISE Winter School in will
feature a range of <a
href="pageX">IR lectures</a> by
experts from the field.
```

"winter school"

"information retrieval"

"IR lectures"

pageX



Fielded Document Representation

```
title:     Winter School 2013
meta:      PROMISE, school, PhD, IR, DB, [...]
           PROMISE Winter School 2013, [...]
headings:  PROMISE Winter School 2013
           Bridging between Information Retrieval and Databases
           Bressanone, Italy 4 - 8 February 2013
body:      The aim of the PROMISE Winter School 2013 on "Bridging between
           Information Retrieval and Databases" is to give participants a
           grounding in the core topics that constitute the multidisciplinary
           area of information access and retrieval to unstructured,
           semistructured, and structured information. The school is a week-
           long event consisting of guest lectures from invited speakers who
           are recognized experts in the field. [...]
anchors:   winter school
           information retrieval
           IR lectures
```

Anchor text is added as a separate document field

Fielded Extension of Retrieval Models

- BM25 => BM25F
- LM => Mixture of Language Models (MLM)

BM25F

- Extension of BM25 incorporating multiple fields
- The soft normalization and term frequencies need to be adjusted
- Original BM25:

$$score(d,q) = \sum_{t\in q} \frac{f_{t,d}\cdot(1+k_1)}{f_{t,d}+k_1\cdot B}\cdot idf_t$$

where B is the soft normalization:

$$B = (1-b+b\frac{|d|}{avgdl})$$

BM25F

$$score(d,q) = \sum_{t\in q} \frac{\tilde{f}_{t,d}}{k_1 + \underbrace{\tilde{f}_{t,d}}} \cdot idf_t$$

Combining term frequencies across fields

$$\tilde{f}_{t,d} = \sum_i \underbrace{w_i}_{\text{Field weight}} \underbrace{\frac{f_{t,d_i}}{B_i}}_{\text{Soft normalization for field } i}$$

Field weight

Soft normalization for field *i*

$$B_i = (1 - \underbrace{b_i} + b_i \frac{|d_i|}{avgdl_i})$$

Parameter b becomes field-specific

Mixture of Language Models

- Build a separate language model for each field
- Take a linear combination of them

$$P(t|\theta_d) = \sum_i \underbrace{\mu_i} P(t|\underbrace{\theta_{d_i}})$$

Field language model
Smoothed with a collection model built from all document representations of the same type in the collection

Field weights
$$\sum_{j=1}^{m} \mu_j = 1$$

Field Language Model

Smoothing parameter

$$P(t|\theta_{d_i}) = (1-\lambda_i)\underbrace{P(t|d_i)}_{\text{Empirical field model}} + \lambda_i \underbrace{P(t|C_i)}_{\text{Collection field model}}$$

Empirical field model

Maximum likelihood estimates

Collection field model

$$\frac{f_{t,d_i}}{|d_i|} \qquad \frac{\sum_{d'} f_{t,d'_i}}{\sum_{d'} |d'_i|}$$

Example

$$q = \{\text{IR}, \text{winter}, \text{school}\}$$
$$\text{fields} = \{\text{title}, \text{meta}, \text{headings}, \text{body}\}$$
$$\mu = \{0.2, 0.1, 0.2, 0.5\}$$

$$P(q|\theta_d) = \underbrace{P(\text{“IR”}|\theta_d)} \cdot P(\text{“winter”}|\theta_d) \cdot P(\text{“school”}|\theta_d)$$

$$\begin{aligned} P(\text{“IR”}|\theta_d) &= 0.2 \cdot P(\text{“IR”}|\theta_{d_{title}}) \\ &+ 0.1 \cdot P(\text{“IR”}|\theta_{d_{meta}}) \\ &+ 0.2 \cdot P(\text{“IR”}|\theta_{d_{headings}}) \\ &+ 0.5 \cdot P(\text{“IR”}|\theta_{d_{body}}) \end{aligned}$$

Parameter Estimation for Fielded Language Models

- Smoothing parameter
  - Dirichlet smoothing with avg. representation length
- Field weights
  - Heuristically (e.g., proportional to the length of text content in that field)
  - Empirically (using training queries)
    - Extensive parameter sweep
      - Computationally intractable for more than a few fields

Exercise

Document Importance

Motivation

- There are *query-independent* factors determining a documents’ importance
  - Recency
  - Credibility
  - SPAM
  - …

Incorporating Document Importance

- Typically a static score, computed at indexing time to influence the ranking
  - Sometimes called "boost factor"

$$score'(d,q) = \underbrace{score(d)}_{\substack{\textbf{Query-independent score} \\ \text{"Static" document score}}} \cdot \underbrace{score(d,q)}_{\substack{\textbf{Query-dependent score} \\ \text{"Dynamic" document score}}}$$

Using Language Models

- Language models offer a theoretically sound way of incorporating document importance through *document priors*

$$P(d|q) = \frac{P(q|d)P(d)}{P(q)} \propto P(q|d)\underbrace{\boxed{P(d)}}_{\textbf{Document prior}}$$

- Computation in the log space:

$$\log P(d|q) \propto \log P(q|d) + \log P(d)$$

Parameter Settings

Setting Parameter Values

- Retrieval models often contain parameters that must be tuned to get best performance for specific types of data and queries
- For experiments:
  - Use *training* and *test* data sets
  - If less data available, use *cross-validation* by partitioning the data into *K* subsets

Finding Parameter Values

- Many techniques used to find optimal parameter values given training data
  - Standard problem in machine learning
- In IR, often explore the space of possible parameter values by *grid search ("brute force")*
  - Perform a sweep over the possible parameter values of each parameter, e.g., from 0 to 1 in 0.1 steps