Feature Selection

- Bag of words — reasonable model (at least for "short" messages).

Feature Selection

- Bag of words — reasonable model (at least for "short" messages).

- Term ranking — fast computation (stream of messages).

Feature Selection

- Bag of words — reasonable model (at least for "short" messages).
- Term ranking — fast computation (stream of messages).
- Small sets of **words** — comprehensible output.

Term Ranking

	Doc	Terms			
		cat	*dog*	*beer*	*milk*
$\mathbb{R}$el	**A**	**1**	**1**	**0**	**0**
	B	**1**	**0**	**0**	**1**
	C	**0**	**1**	**1**	**0**
$\mathbb{I}$rr	**U**	**0**	**1**	**0**	**1**
	V	**1**	**0**	**1**	**0**

Term Ranking

	Doc	Terms			
		cat	*dog*	*beer*	*milk*
$\mathbb{R}$el	**A**	**1**	**1**	**0**	**0**
	B	**1**	**0**	**0**	**1**
	C	**0**	**1**	**1**	**0**
$\mathbb{I}$rr	**U**	**0**	**1**	**0**	**1**
	V	**1**	**0**	**1**	**0**

$\mathbf{a}(cat) = \mathbf{2}$

Term Ranking

	Doc	Terms			
		cat	*dog*	*beer*	*milk*
$\mathbb{R}$el	**A**	**1**	**1**	**0**	**0**
	B	**1**	**0**	**0**	**1**
	C	**0**	**1**	**1**	**0**
$\mathbb{I}$rr	**U**	**0**	**1**	**0**	**1**
	V	**1**	**0**	**1**	**0**

$\mathbf{a}(cat) = \mathbf{2}$

$\mathbf{b}(cat) = \mathbf{1}$

Term Ranking

	Doc	Terms			
		cat	*dog*	*beer*	*milk*
$\mathbb{R}$el	**A**	**1**	**1**	**0**	**0**
	B	**1**	**0**	**0**	**1**
	C	**0**	**1**	**1**	**0**
$\mathbb{I}$rr	**U**	**0**	**1**	**0**	**1**
	V	**1**	**0**	**1**	**0**

a(*cat*) = **2**

b(*cat*) = **1**

Hamming: $\mu(t) = \mu(a(t), b(t)) = \frac{a(t)|\mathbb{I}| + b(t)|\mathbb{R}| - 2a(t)b(t)}{|\mathbb{R}||\mathbb{I}|}$

Term Ranking

	Doc	Terms			
		cat	*dog*	*beer*	*milk*
$\mathbb{R}$el	**A**	**1**	**1**	**0**	**0**
	B	**1**	**0**	**0**	**1**
	C	**0**	**1**	**1**	**0**
$\mathbb{I}$rr	**U**	**0**	**1**	**0**	**1**
	V	**1**	**0**	**1**	**0**

a(*cat*) = **2**

b(*cat*) = **1**

Hamming: $\mu(t) = \mu(a(t), b(t)) = \frac{a(t)|\mathbb{I}| + b(t)|\mathbb{R}| - 2a(t)b(t)}{|\mathbb{R}||\mathbb{I}|}$

Correlation: $\gamma(t) = \gamma(a(t), b(t)) = \frac{\left|\frac{a(t)}{\mathbb{R}} - \frac{b(t)}{\mathbb{I}}\right|}{(a(t)+b(t))(|\mathbb{R}|+|\mathbb{I}|-a(t)-b(t))}$

Term Ranking

	Doc	Terms			
		cat	*dog*	*beer*	*milk*
$\mathbb{R}$el	**A**	**1**	**1**	**0**	**0**
	B	**1**	**0**	**0**	**1**
	C	**0**	**1**	**1**	**0**
$\mathbb{I}$rr	**U**	**0**	**1**	**0**	**1**
	V	**1**	**0**	**1**	**0**

a(*cat*) = **2**

b(*cat*) = **1**

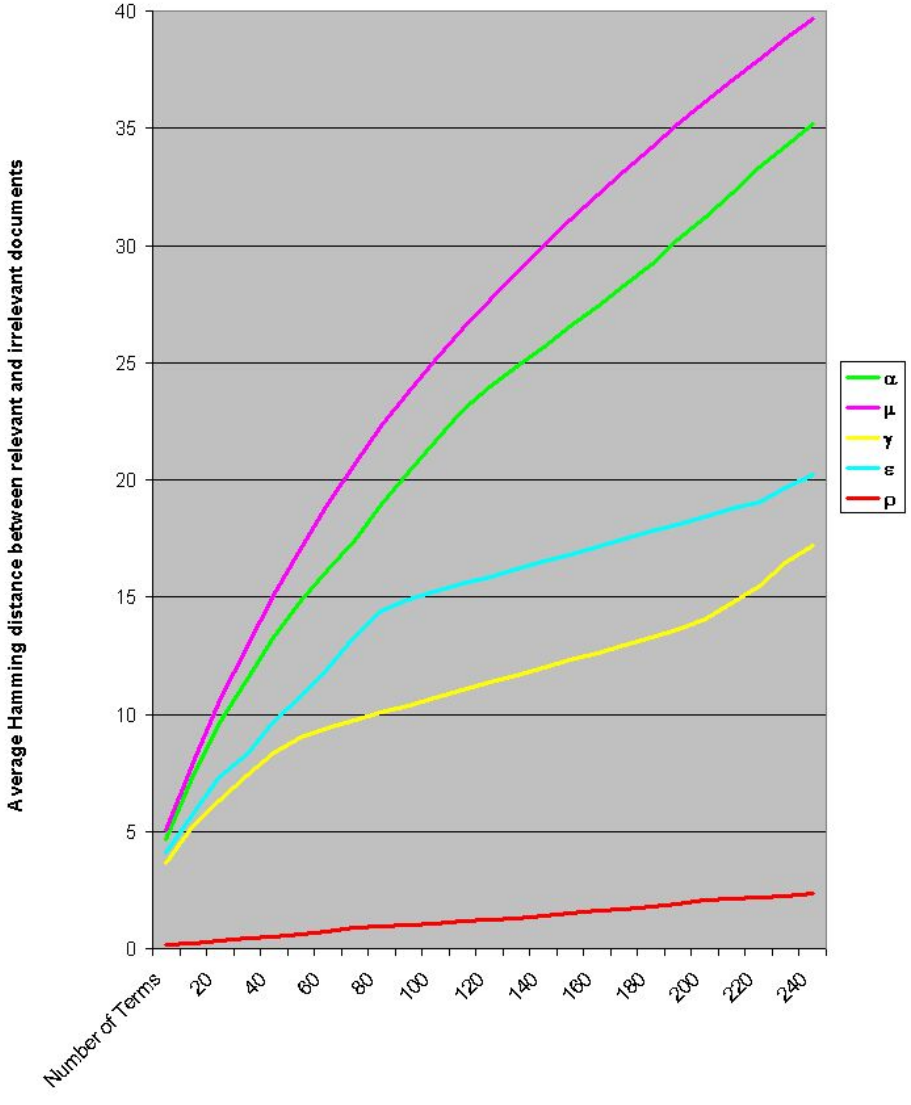
Hamming: $\mu(t) = \mu(a(t), b(t)) = \frac{a(t)|\mathbb{I}| + b(t)|\mathbb{R}| - 2a(t)b(t)}{|\mathbb{R}||\mathbb{I}|}$

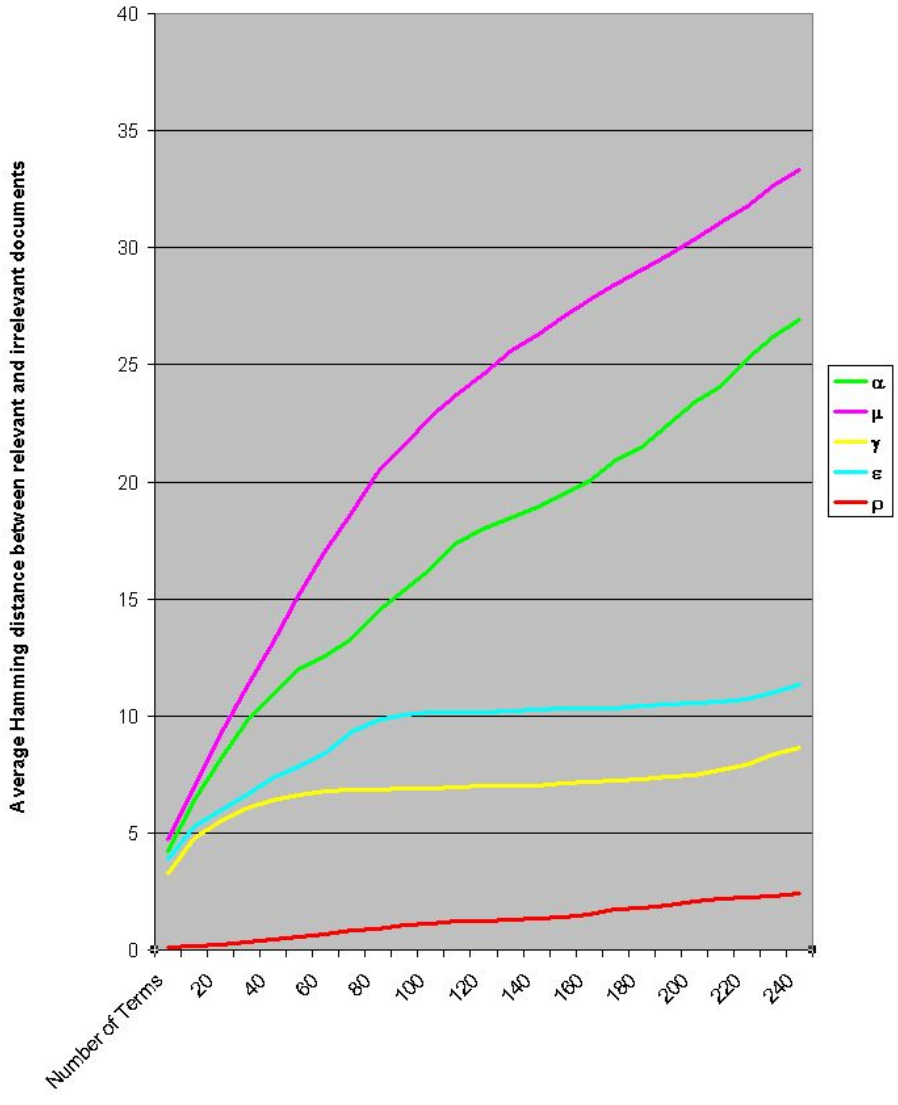
Correlation: $\gamma(t) = \gamma(a(t), b(t)) = \frac{\left|\frac{a(t)}{\mathbb{R}} - \frac{b(t)}{\mathbb{I}}\right|}{(a(t)+b(t))(|\mathbb{R}|+|\mathbb{I}|-a(t)-b(t))}$

Entropy: $\epsilon(t) = \epsilon(a(t), b(t)) = \quad \ldots$

$\ldots$

Separation Results





Validation: TREC 2002 Batch Filtering

Approach

Choose $|T| = 50$ top words by majority vote of 5 best measures.

Retrieve test document d if $\tau < \sum_{t \in d \cap T} \log \left(\frac{a(t)+0.5}{|\mathbb{R}|+1} / \frac{b(t)+0.5}{|\mathbb{I}|+1} \right)$

Optimize τ for utility $TU11 = 2R - I$ on the training set.

Validation: TREC 2002 Batch Filtering

Approach

Choose $|T| = 50$ top words by majority vote of 5 best measures.

Retrieve test document d if $\tau < \sum_{t \in d \cap T} \log \left(\frac{a(t)+0.5}{|\mathbb{R}|+1} / \frac{b(t)+0.5}{|\mathbb{I}|+1} \right)$

Optimize τ for utility $TU11 = 2R - I$ on the training set.

Results

On 100 topics our TU11 score was

16 times below median,
26 times equal to the median and
58 times above median (21 times equal to the maximum).

Our system ranked 6th out of the 15 competing systems.

Future Research

- Bag of words vs. bag of n-gramms

Future Research

- Bag of words vs. bag of n-gramms

- Event detection in message streams based on the slope of average Hamming distance curve ...

Future Research

- Bag of words vs. bag of n-gramms
- Event detection in message streams based on the slope of average Hamming distance curve ...
- Concept generation ...