Putting Suffix-Tree-Stemming to Work

Benno Stein
Bauhaus University Weimar

Martin Potthast
Paderborn University

Index terms

Text with markups [Reuters]:

```
<TEXT> <TITLE>CHRYSLER> DEAL LEAVES UNCERTAINTY
FOR AMC WORKERS</TITLE> <AUTHOR> By Richard
Walker, Reuters</AUTHOR> <DATELINE> DETROIT,
March 11 - </DATELINE><BODY>Chrysler Corp’s 1.5
billion dlr bid to takeover American Motors Corp;
AMO> should help bolster the small automaker’s
sales, but it leaves the future of its 19,000
employees in doubt, industry analysts say.  It
was "business as usual" yesterday at the American
...
```

Index terms

Raw text:

```
chrysler deal leaves uncertainty for amc workers
by richard walker reuters detroit march 11
chrysler corp s 1 5 billion dlr bid to takeover
american motors corp should help bolster the
small automaker s sales but it leaves the future
of its 19 000 employees in doubt industry
analysts say it was business as usual yesterday
at the american
```

Index terms

Stop words emphasized:

```
chrysler deal leaves uncertainty for amc workers
by richard walker reuters detroit march 11
chrysler corp s 1 5 billion dlr bid to takeover
american motors corp should help bolster the
small automaker s sales but it leaves the future
of its 19 000 employees in doubt industry
analysts say it was business as usual yesterday
at the american
```

Index terms

After stemming:

`chrysler deal leav uncertain amc work richard walk reut detroit takeover american motor help bols automak sal leav futur employ doubt industr analy business usual yesterday`

Index terms

After stemming:

`chrysler deal leav uncertain amc work richard walk reut detroit takeover american motor help bols automak sal leav futur employ doubt industr analy business usual yesterday`

Stemming algorithms remove inflectional and morphological affixes.

connect	connects
	connected
	connecting
	connection

Index terms

After stemming:

```
chrysler deal leav uncertain amc work richard
walk reut detroit takeover american motor help
bols automak sal leav futur employ doubt industr
analy business usual yesterday
```

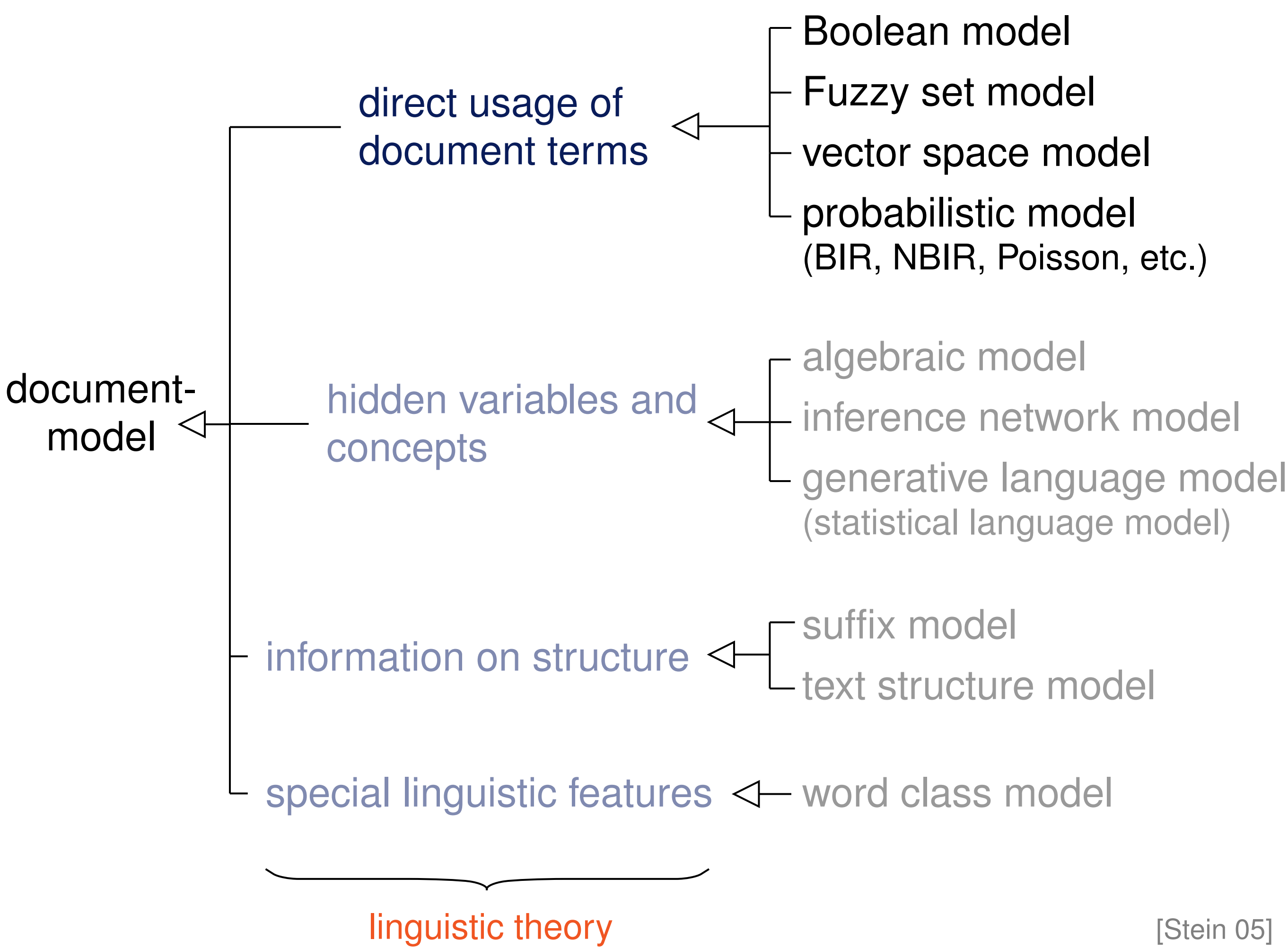
Stemming algorithms remove inflectional and morphological affixes.

```
connect      connects
             connected
             connecting
             connection
```

+ make text operations less dependent on special word forms
+ reduce the dictionary size

– may merge words that have very different meanings
– discard possibly useful information about language use

Index terms



[Stein 05]

Retrieval model $\sim$ document model

Stemming Approaches

1. Table lookup.
   To each stem all flections are stored in a hash table.
   Problem: memory size (consider client-side applications)

2. Successor variety analysis.
   Morpheme boundaries are found by statistical analyses.
   Problem: parameter settings, runtime

3. Affix elimination.
   Rule-based replacement of prefixes and suffixes;
   the most commonly used approach.

   Principle: *iterative longest match stemming*

   (a) Removal of the match resulting from the longest precondition.
   (b) Exhaustive application of the first step.
   (c) Repair of irregularities.

Stemming Approaches

Affix Elimination under Porter

Rule type	Condition	Suffix	Replacement	Example
1a	Null	sses	ss	caresses $\rightarrow$ caress
1a	Null	ies	i	ponies $\rightarrow$ poni
1b	(m>0)	eed	ee	feed $\rightarrow$ feed agreed $\rightarrow$ agree
1b	(*v*)	ed	ε	plastered $\rightarrow$ plaster bled $\rightarrow$ bled
1b	(*v*)	ing	ε	motoring $\rightarrow$ motor sing $\rightarrow$ sing
1c	(*v*)	y	i	happy $\rightarrow$ happi sky $\rightarrow$ sky
2	(m>0)	biliti	ble	sensibiliti $\rightarrow$ sensible

(m>x)	number of vocal-consonant-sequences exceeds x
(*S)	stem ends with letter S
(*v*)	stem contains vocal
(*o)	stem ends with cvc where second consonant $c \notin \{W, X, Y\}$
(*d)	stem ends with two identical consonants

Stemming Approaches

Affix Elimination under Porter: Weaknesses

- difficult to modify:
  effects of new rules are barely to anticipate

- subject to over-generalization:
  `policy/police university/universe`
  `organization/organ`

- several definite generalizations are not covered:
  `European/Europe matrices/matrix`
  `machine/machinery`

- generates stem that are hard to be interpreted:
  `iteration/iter general/gener`

Stemming Approaches

Successor Variety Analysis: Interesting Aspects

- The idea of *corpus-specific stemming*.
  Corpus dependency is an advantage, if the corpus has a strong topic or application bias.

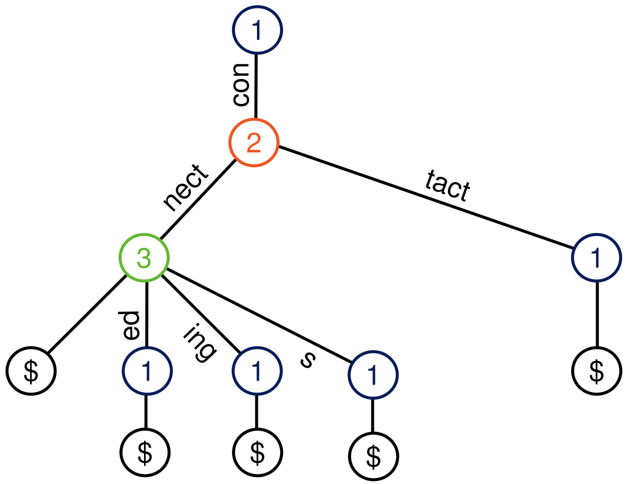
- The idea of *language independence*.
  Language independence is essential for multilingual documents or if the language cannot be determined.

Stemming approach	Corpus dependency	Language independence
Affix elimination	no	yes
Variety analysis	yes	little

Stemming Approaches

Successor Variety Analysis: Realization

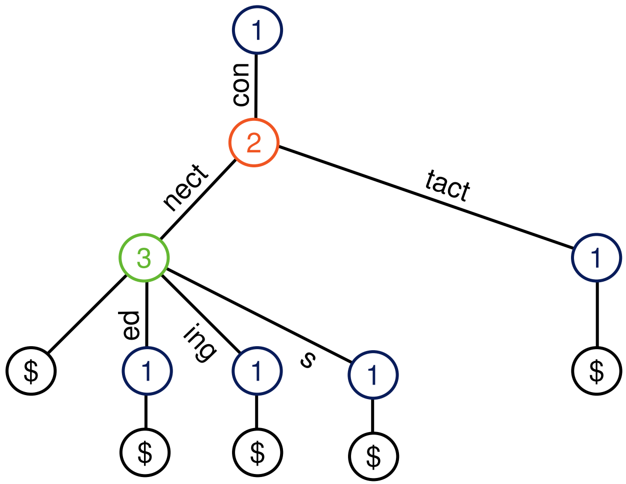
Suffix tree at letter level:



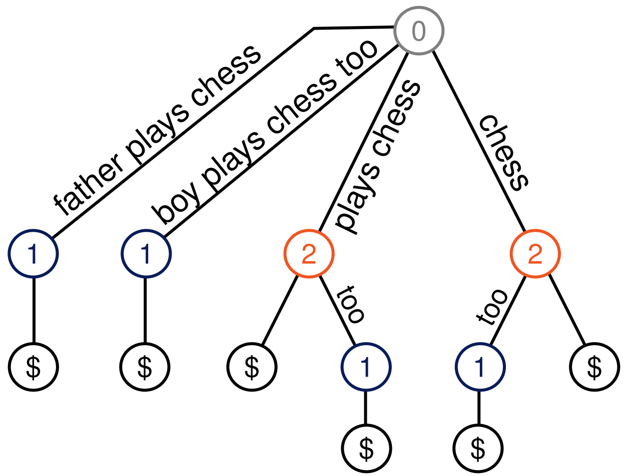
Stemming Approaches

Successor Variety Analysis: Realization

Suffix tree at letter level:



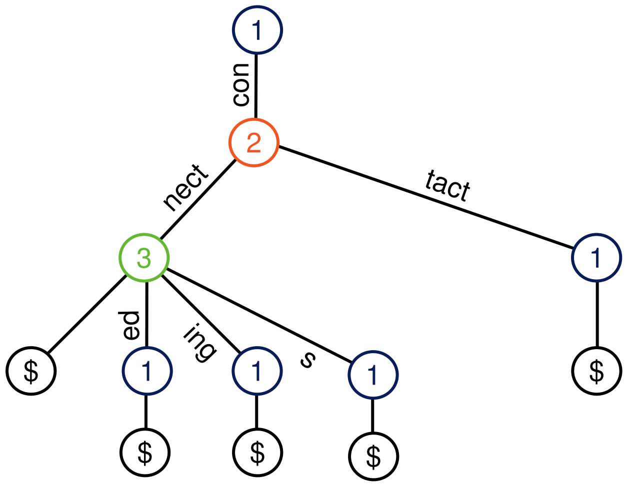
Suffix tree at word level:



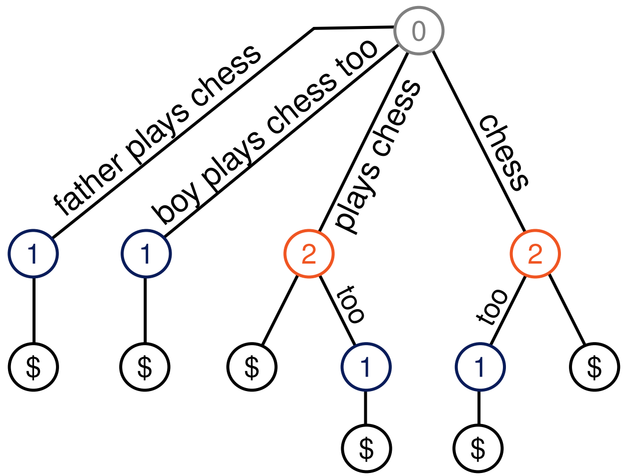
Stemming Approaches

Successor Variety Analysis: Realization

Suffix tree at letter level:



Suffix tree at word level:



How to find good candidates for a stem?

- analysis of degree differences (depending on tree depth)
- cut-off method, complete word method, entropy method

Evaluation

Caution is advised ;)

- existing reports on the impact of stemming are contradictory
- employed analysis tool (among others): clustering

  But what can be found?

  1. improved document model
  2. peculiarity of a clustering algorithm
  3. ...

Evaluation

Caution is advised ;)

- ❑ existing reports on the impact of stemming are contradictory
- ❑ employed analysis tool (among others): clustering

  But what can be found?

  1. improved document model
  2. peculiarity of a clustering algorithm
  3. . . .

A cluster algorithm's performance depends on various parameters.

Different cluster algorithms behave differently sensitive to document model "improvements".

Baseline? Interpretation? Objectivity? Generalizability?

Evaluation

Caution is advised ;)

An objective way to rank document models is to compare their ability to *capture the intrinsic similarity relations* of a collection D.

Basic idea:

1. construct a similarity graph, $G = \langle V, E, w \rangle$
2. measure its conformance to a reference classification
3. analyze improvement/decline under new document model

Expected Density $\bar{\rho}$

Definition

Graph $G = \langle V, E, w \rangle$

- G is called sparse [dense] if $|E| = O(|V|)$ $[O(|V|^2)]$
- the density θ computes from the equation $|E| = |V|^\theta$

Expected Density $\bar{\rho}$

Definition

Graph $G = \langle V, E, w \rangle$

- G is called sparse [dense] if $|E| = O(|V|)$ $[O(|V|^2)]$
- the density θ computes from the equation $|E| = |V|^\theta$
- with $w(G) := \sum_{e \in E} w(e)$, this extends to weighted graphs:

$$w(G) = |V|^\theta \quad \Leftrightarrow \quad \theta = \frac{\ln(w(G))}{\ln(|V|)}$$

Using θ we assess the density of an induced subgraph G_i of G.

Expected Density $\bar{\rho}$

Definition

Graph $G = \langle V, E, w \rangle$

- G is called sparse [dense] if $|E| = O(|V|)$ $[O(|V|^2)]$
- the density θ computes from the equation $|E| = |V|^\theta$
- with $w(G) := \sum_{e \in E} w(e)$, this extends to weighted graphs:

$$w(G) = |V|^\theta \quad \Leftrightarrow \quad \theta = \frac{\ln(w(G))}{\ln(|V|)}$$

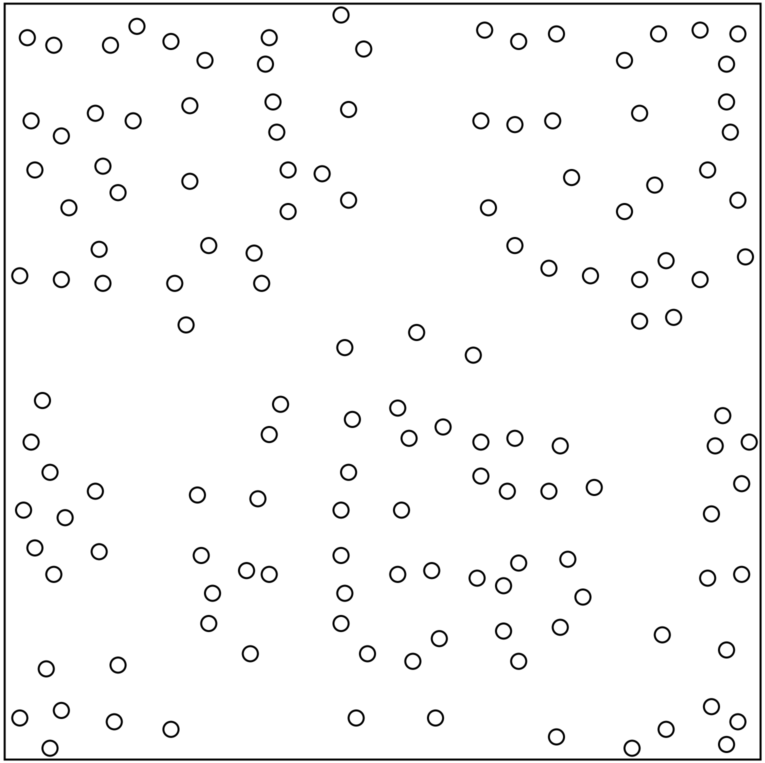
Using θ we assess the density of an induced subgraph G_i of G.

- a categorization $\mathcal{C} = \{C_1, \ldots, C_k\}$ induces k subgraphs G_i

→ expected density $\overline{\rho}(\mathcal{C}) = \sum_{i=1}^{k} \frac{|V_i|}{|V|} \cdot \frac{w(G_i)}{|V_i|^\theta}$

Expected Density $\bar{\rho}$

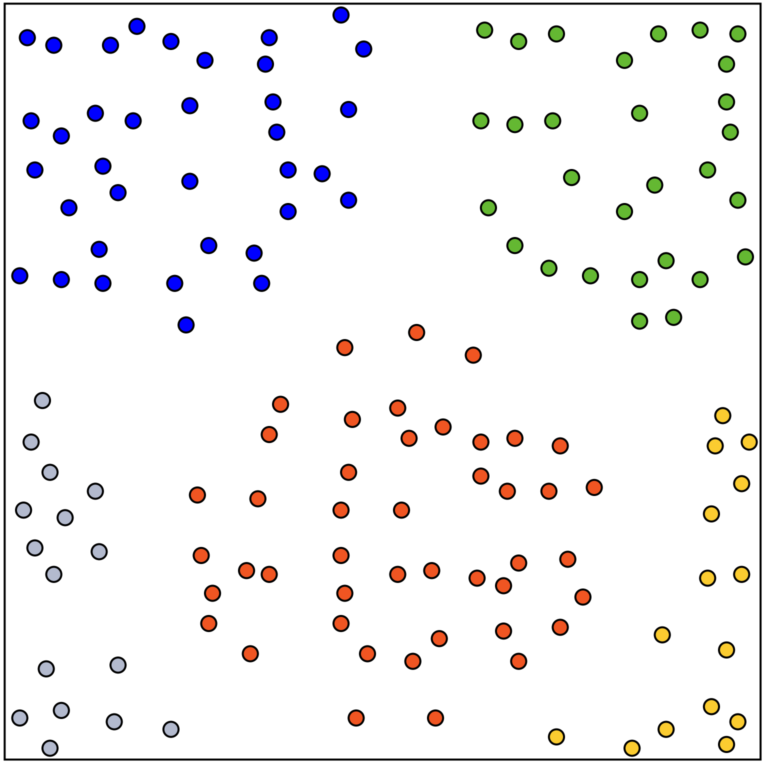
Understanding Expected Density



Embedding of a collection under a particular document model.

Expected Density $\bar{\rho}$

Understanding Expected Density

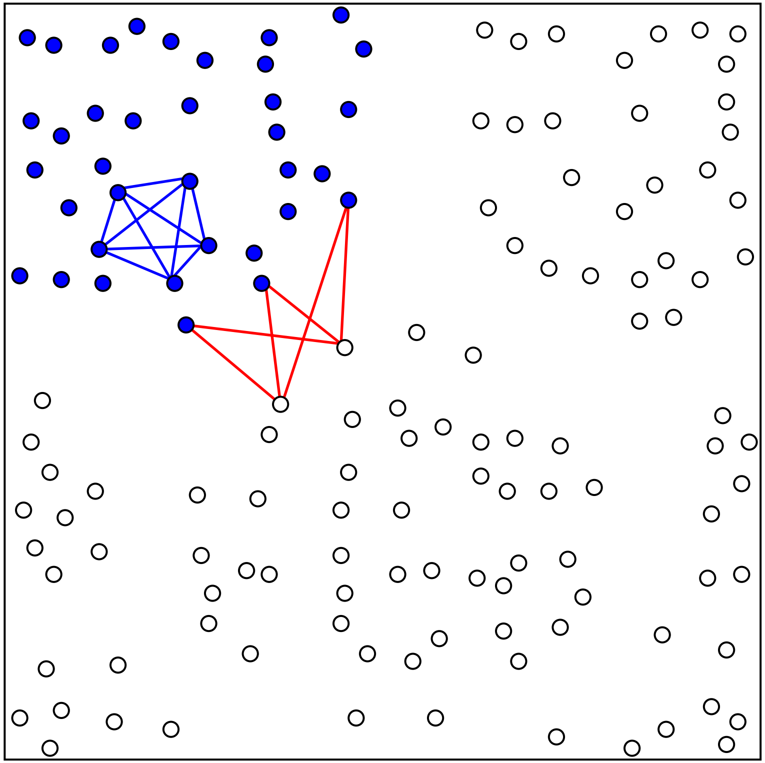


Embedding of a collection under a particular document model.

$\overline{\rho} > 1$ [$\overline{\rho} < 1$] if the cluster density is larger [smaller] than average.

Expected Density $\bar{\rho}$

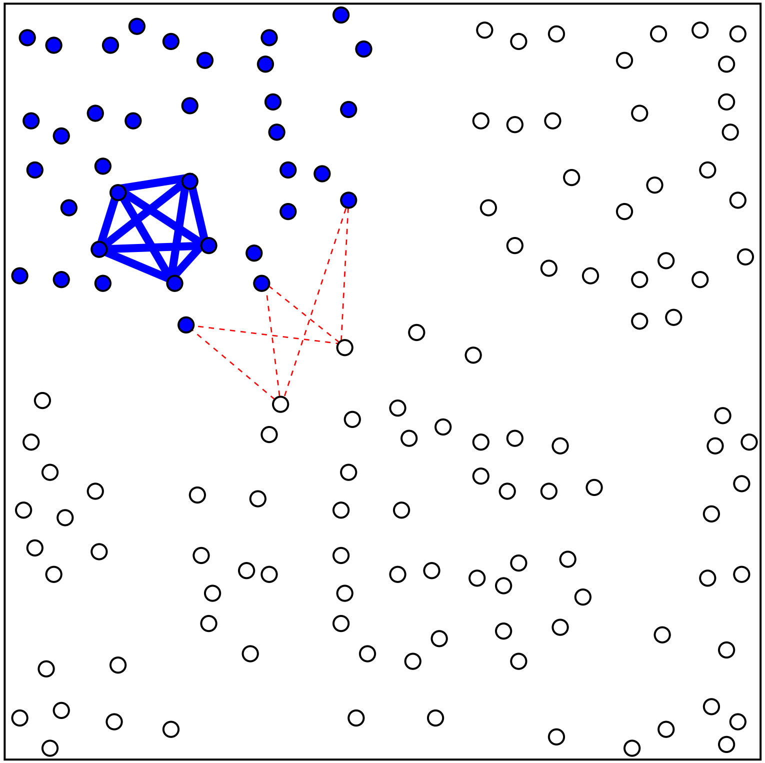
Understanding Expected Density



Consider inter-cluster and intra-cluster similarities.

Expected Density $\bar{\rho}$

Understanding Expected Density

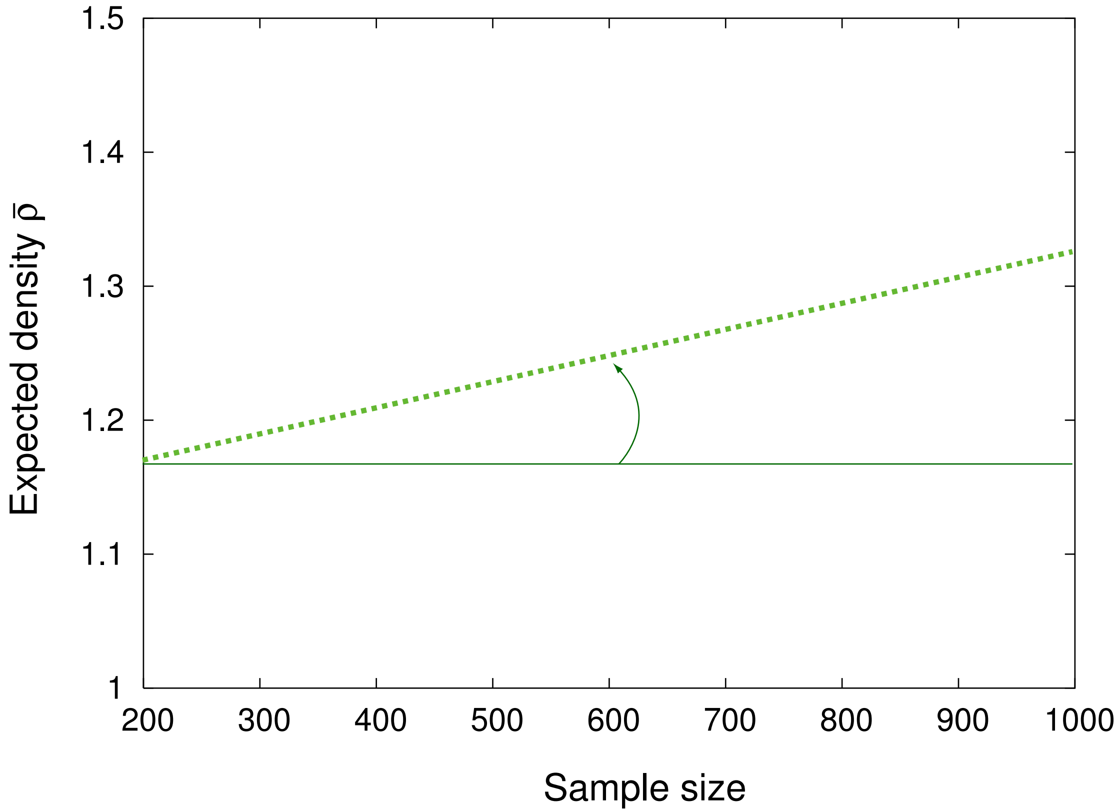


Consider inter-cluster and intra-cluster similarities.

Effect of a document model that *reinforces the structural characteristic* within a document collection.

Expected Density $\bar{\rho}$

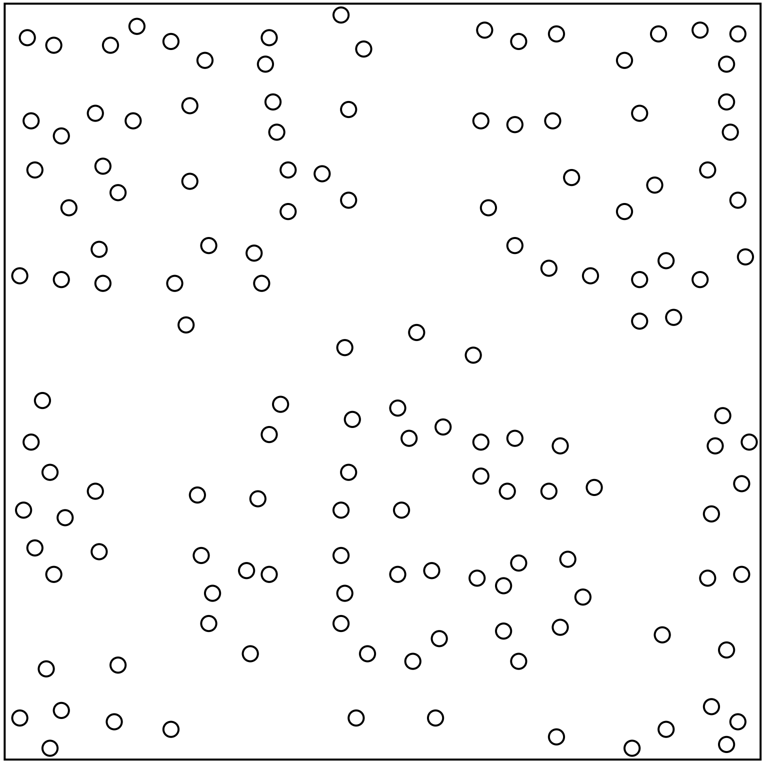
Understanding Expected Density



The expected density $\bar{\rho}$ is a monotonically increasing function of the sample size.

Expected Density $\bar{\rho}$

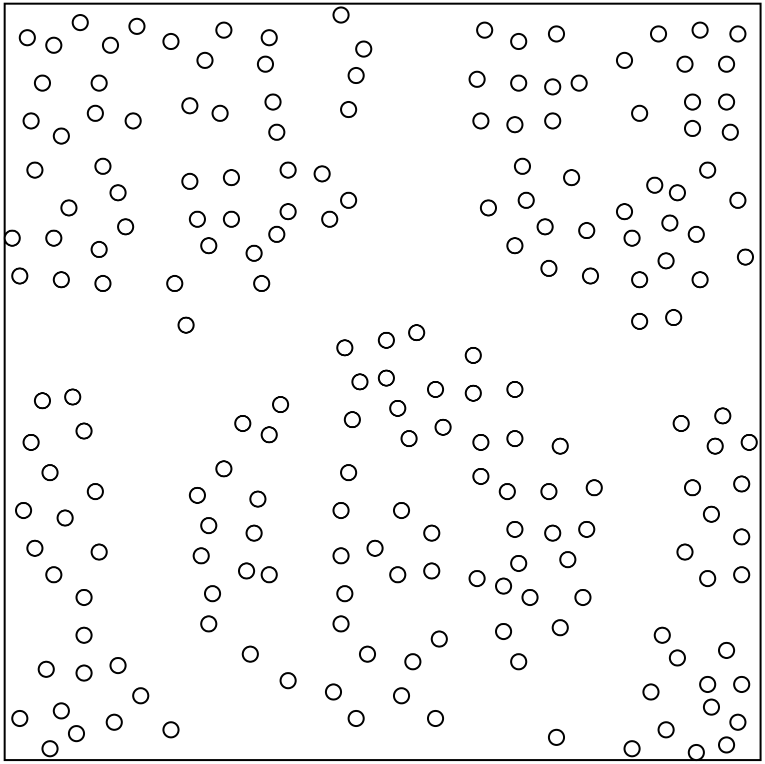
Understanding Expected Density



The expected density $\overline{\rho}$ is a monotonically increasing function of the sample size.

Expected Density $\bar{\rho}$

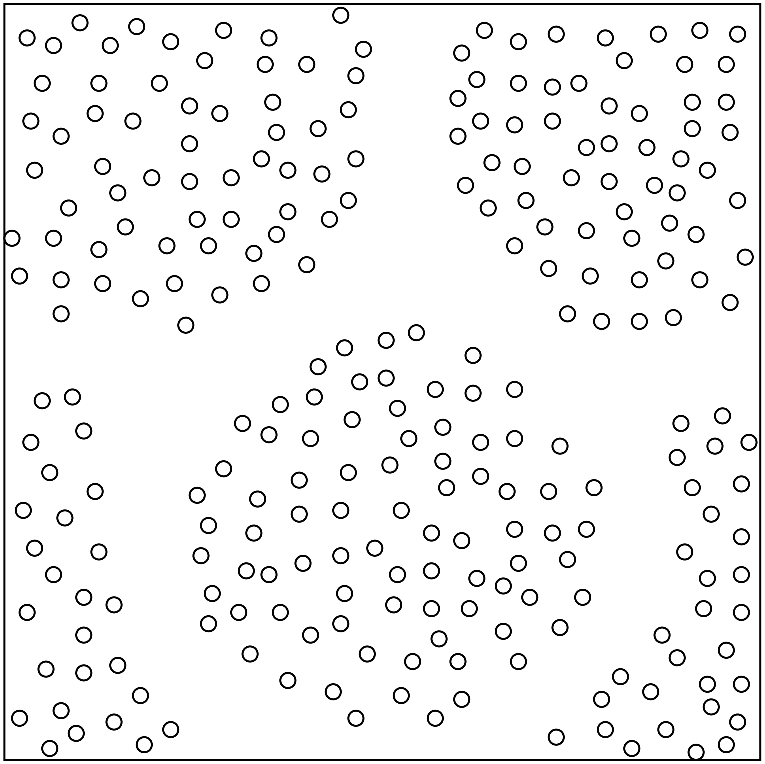
Understanding Expected Density



The expected density $\overline{\rho}$ is a monotonically increasing function of the sample size.

Expected Density $\bar{\rho}$

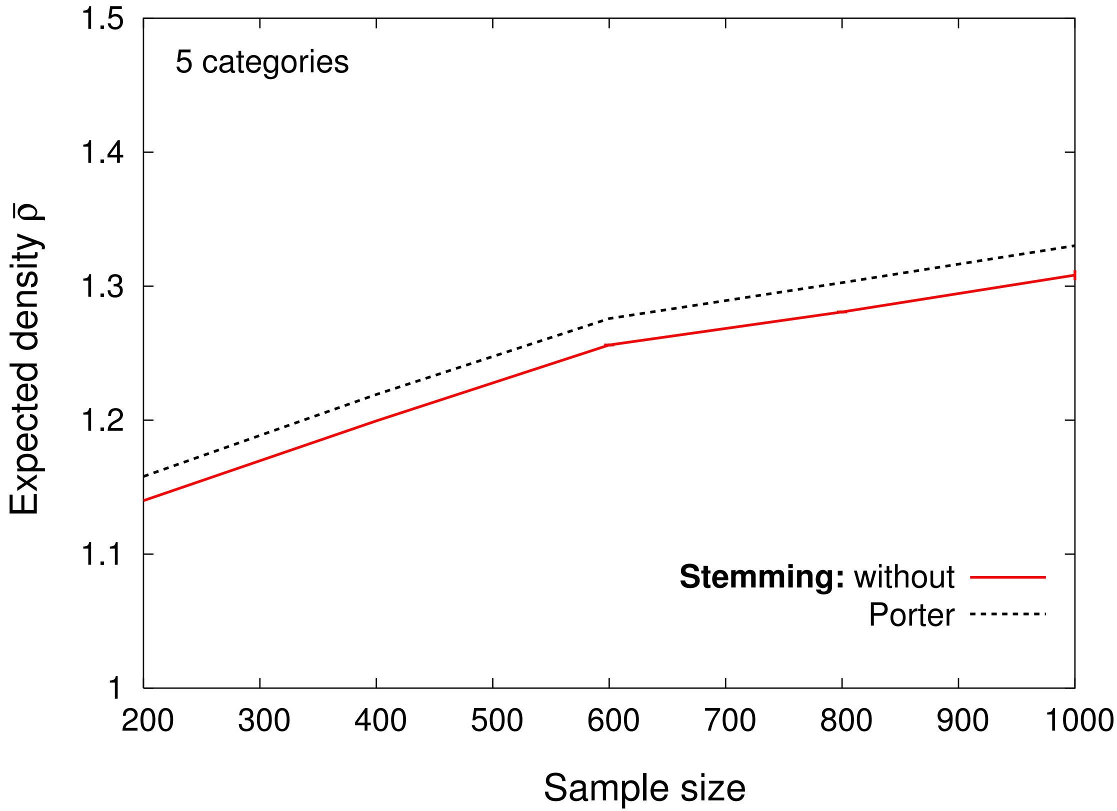
Understanding Expected Density



The expected density $\overline{\rho}$ is a monotonically increasing function of the sample size.

Expected Density $\bar{\rho}$

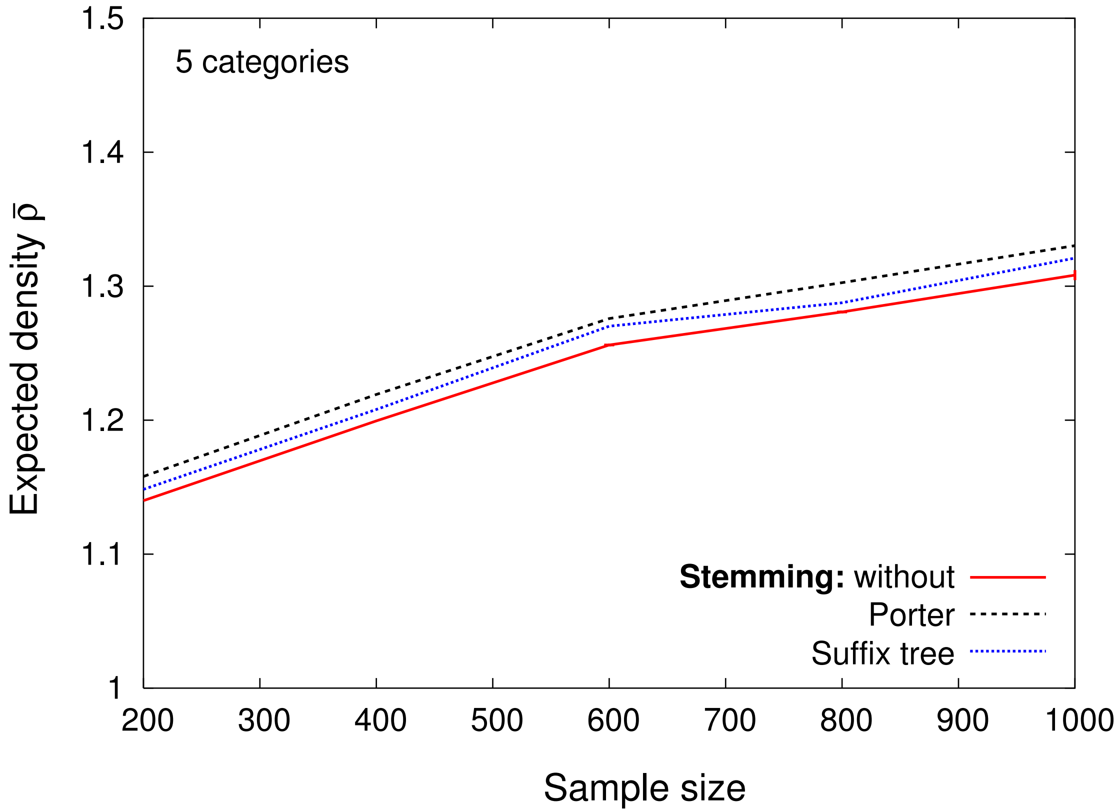
Experiments: English Collection



Collection: RCV1. Two documents d_1, d_2 are assigned to the same category if they share the top level category and the most specific category.

Expected Density $\bar{\rho}$

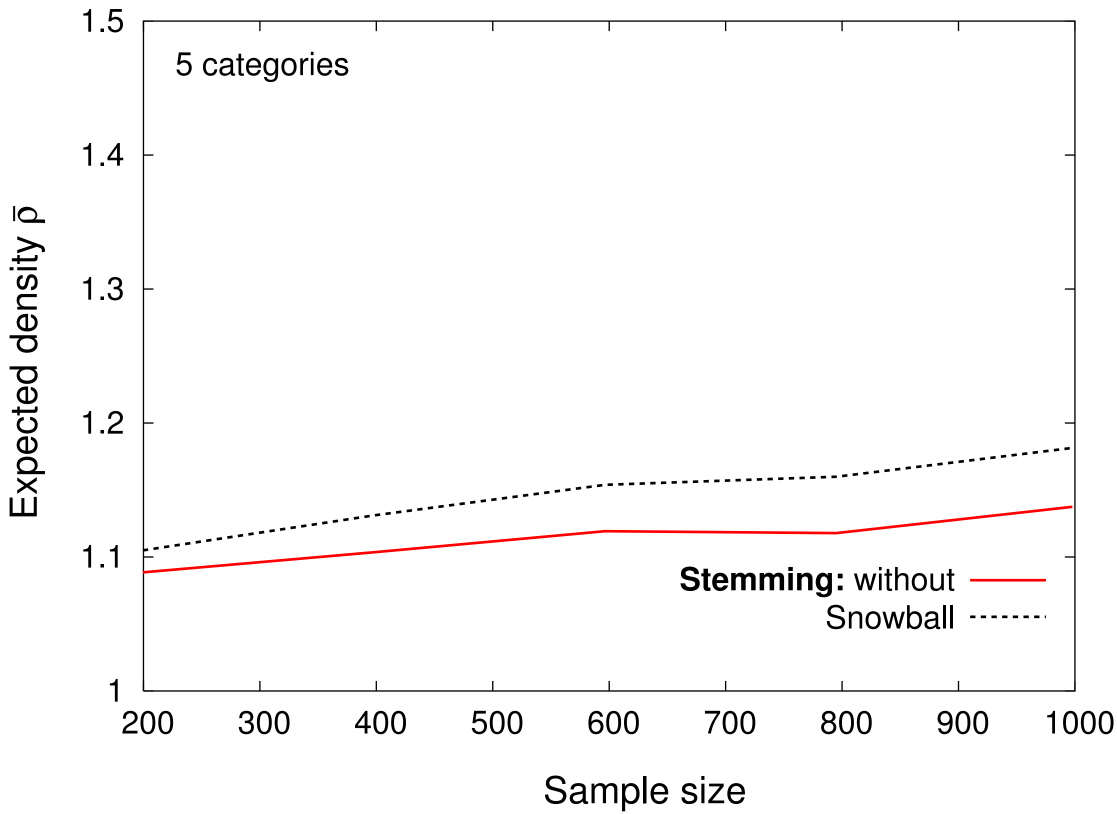
Experiments: English Collection



A note on reproducibility: meta information files that describe the compiled test collections are made available upon request.

Expected Density $\bar{\rho}$

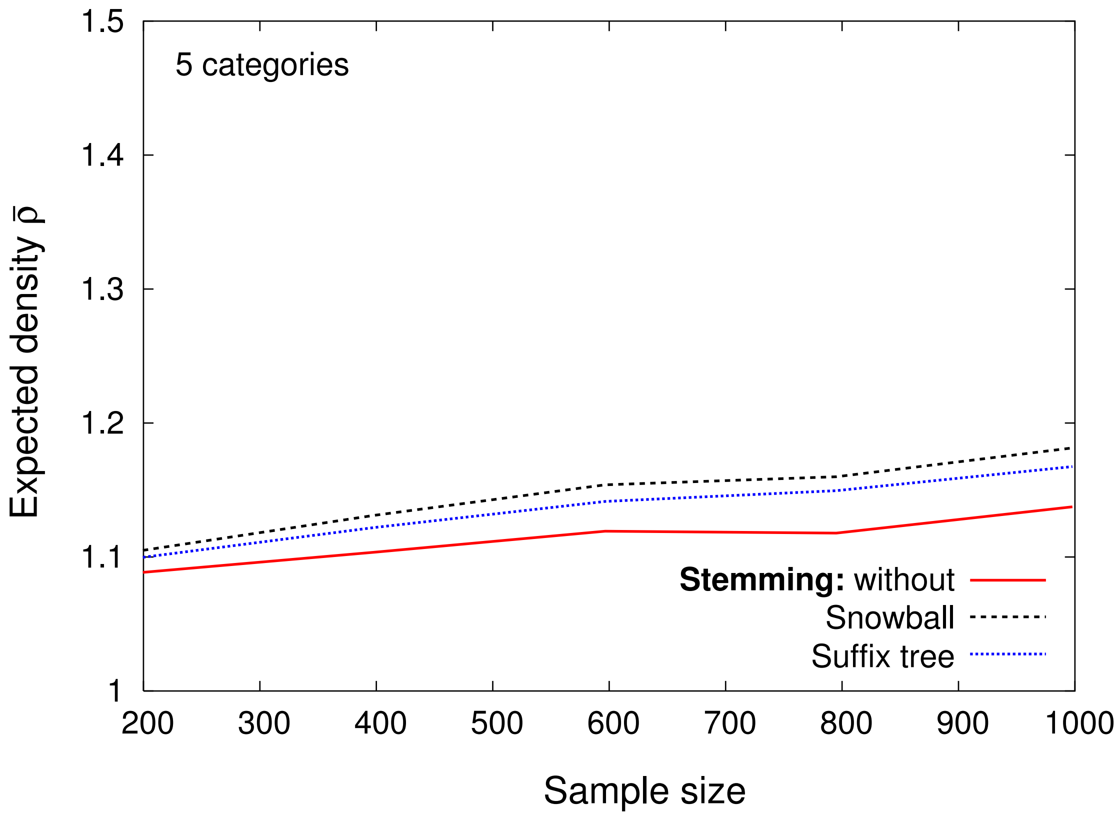
Experiments: German Collection



Collection: Compilation of 26,000 documents from 20 German news groups.

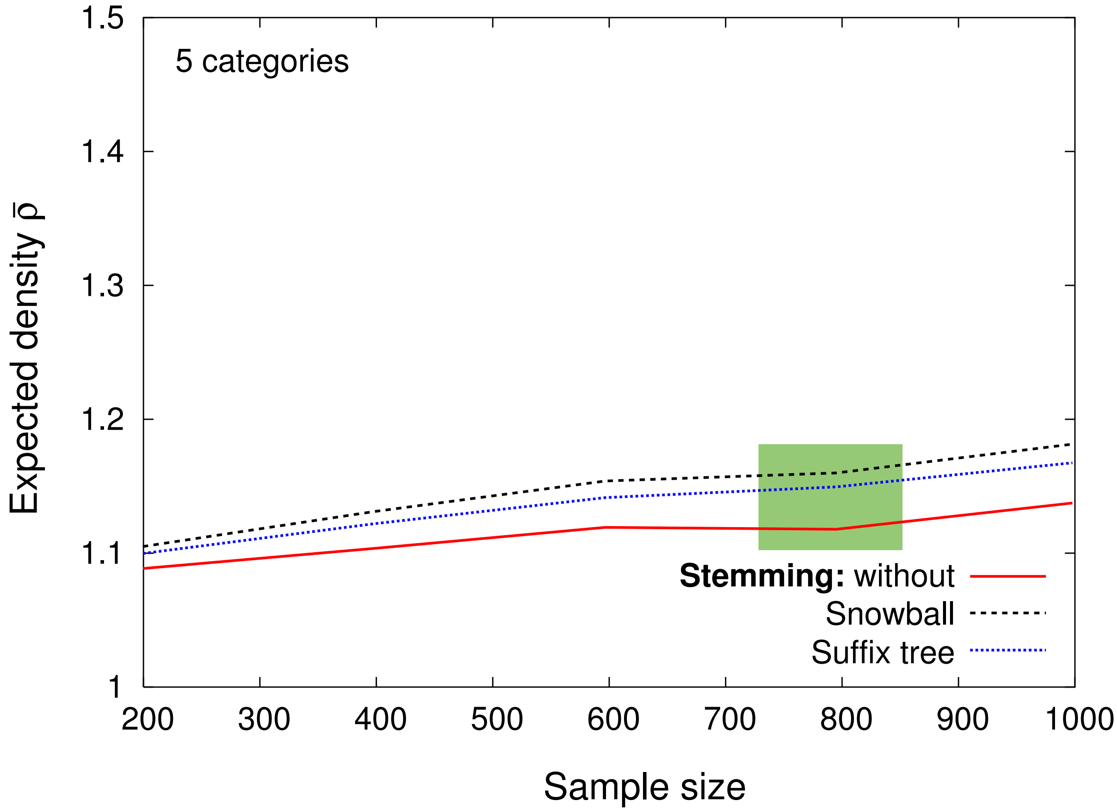
Expected Density $\bar{\rho}$

Experiments: German Collection



Expected Density $\bar{\rho}$

Experiments: German Collection



Stemming can reduce noise.

Expected Density $\bar{\rho}$

Experiments: German Collection

Where successor variety works:

mechanis	–	mus, tisch, che, ch, tischen, men,
	–	tisches, ierung, chen
zusammen	–	leben, gang, h
zusammenbr	–	icht, uch, aut, echen
zusammenfass	–	en, ung, t, end
zusammenge	–	faßt, baut, zählt, fasst
zusammengesetzt	–	en, $
zusammenh	–	ängen, ängt, änge
zusammenha	–	lten, lt
zusammenhang	–	los, es, s, $

Expected Density $\bar{\rho}$

Experiments: German Collection

Where successor variety works:

`mechanis`	–	`mus, tisch, che, ch, tischen, men,`
	–	`tisches, ierung, chen`
`zusammen`	–	`leben, gang, h`
`zusammenbr`	–	`icht, uch, aut, echen`
`zusammenfass`	–	`en, ung, t, end`
`zusammenge`	–	`faßt, baut, zählt, fasst`
`zusammengesetzt`	–	`en, $`
`zusammenh`	–	`ängen, ängt, änge`
`zusammenha`	–	`lten, lt`
`zusammenhang`	–	`los, es, s, $`

and where it fails:

`schwarz`	–	`arbeit, denker, schild, fahrer, em, en,`
	–	`e, markt, maler, bader, hörer, radler, e, s`

Evaluation

A Note on F-Measure Values

Stemming approach	F-min	F-max	F-av.
	(sample size 1000, 10 categories)		
without		—baseline—	
Porter	-12%	11%	2%
suffix tree	-10%	10%	2%

Evaluation

A Note on F-Measure Values

Stemming approach	F-min	F-max	F-av.
	(sample size 1000, 10 categories)		
without		—baseline—	
Porter	-12%	11%	2%
suffix tree	-10%	10%	2%

A Note on Runtime

- successor variety analysis with suffix trees
  in $O(n)$ [Ukkonen 1995], and
  in $O(n^2)$ and $\Theta(n \log(n))$ respectively [Giegerich et. al.]
- successor variety analysis with Pat trees
  in $O(n^2)$; $\Theta(n \log(n))$ may be assumed for short affixes

Summary

- Basis: document models with “visible” index terms
- Issue: selection, modification, enrichment of index terms
- Question: stemming without semantic background

Contribution

- efficient implementation of variational stemming with Patricia
- parameter optimization $\Rightarrow$ significantly better than [Frakes 1992]
- comparison to Porter stemmer and Snowball stemmer
- algorithm-neutral evaluation method based on $\bar{\rho}$

Message

- the impact of stemming may be over-estimated
- generally accepted analysis methods are required

Summary

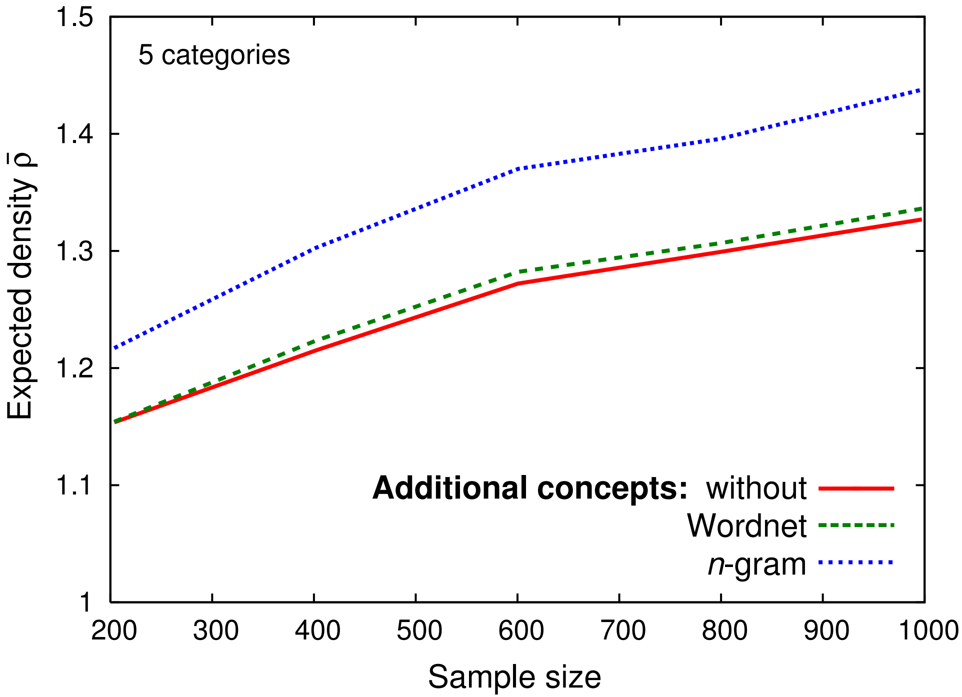
Related Work

- A similar approach can be applied to index construction.
  variational n-grams: use words (not letters) as tokens
- Issue: *collection-specific* document model
- Motto: “co-occurrence analysis versus Wordnet”

Summary

Related Work

- A similar approach can be applied to index construction.
  variational n-grams: use words (not letters) as tokens
- Issue: *collection-specific* document model
- Motto: “co-occurrence analysis versus Wordnet”



References

- Bacchin, M. Ferro, N. and Melucci, M. (2002): Experiments to evaluate a statistical stemming algorithm. *CLEF: Cross-Language Evaluation Forum Workshop*, Rome, 161–168.
- Frakes, W. B. (1984): Term conflation for information retrieval. In: *Proceedings of SIGIR '84*, Swinton, UK, 383–389.
- Fürnkranz, J. (1998): A Study Using n-gram Features for Text Categorization. Austrian Institute for Artificial Intelligence.
- Mayfield, J. and McNamee, P. (2003): Single n-gram stemming. In: *Proceedings of the SIGIR '03*, Toronto, 415–416.
- Porter, M. F. (1980): An Algorithm for Suffix Stripping. *Program*, 14(3):130–137.