Ruprecht-Karls Universität Heidelberg
Seminar für Computerlinguistik
Software Projekt WS 2009/10

Language Identification XXL

Referenten: Galina Sigwarth, Maria Semenchuk, George Kakish, Ulzhan Kadirbayeva
Betreuer: Nils Reiter

Inhalt

- Motivation
- Ziel & Projekaufgabe
- Ablauf
- Stand der Forschung
- Ansätze
- Ressourcen
- Evaluation
- Planung – Aufgabeverteilung
- Quellen

Motivation

- Sprachidentifizierung:

  Ermittlung der Sprache, in der ein elektronisches Textdokument verfasst ist.

- Wozu braucht man Sprachidentifizierungssysteme?
  - Information Retrieval
  - Pre-Processing für andere Techniken der Sprachverarbeitung
  - Rechtschreibkorrektur
  - Optical Character Recognition (OCR)

Unë flas Shqip	Albanian
አማርኛ፡ እችላለሁ።	Amharic
انا اتكلم اللغة العربية	Arabic
Ես Հայերէն կը խօսիմ	Armenian
Мән азәрбајан дилиндә данышырам	Azeri
আমি বাংলা ভাষায় কথা বলি	Bengali
Govorim bosanski/hrvatski	Bosnian/Croatian
Аз говоря български	Bulgarian
ကျွန်ုပ် မြန်မာလိုတတ်ပါသည်။	Burmese
我說粵語	Cantonese
Mluvím česky	Czech
I speak English	English
Ma räägin Eesti keelt	Estonian
من فارسی حرف میزنم	Farsi
Je parle français	French
მე ვლაპარაკობ ქართულად	Georgian
Ich spreche Deutsch	German
હું ગુજરાતી બોલું છું	Gujerati
NA YIA HAUSA	Hausa
אני דובר עברית	Hebrew
मैं हिन्दी बोलता हूँ	Hindi
Beszélek Magyarul	Hungarian
Anam asu igbo	Ibo
Saya bicara bahasa	Indonesian
Мен казақша билемин	Kazakh
Nvuga ikinyarwanda	Kinyarwanda
나는 한국말을 합니다	Korean
من بەکوردی قسە دەکەم	Kurdish
Es runāju latviski	Latvian
Na lobaka Lingala	Lingala
Aš kalbu lietuviškai	Lithuanian
Јас зборувам македонски	Macedonian
Saya bicara bahasa Malay	Malay
我说汉语	Mandarin
मी मराठी बोलतो	Marathi
Би Монгол хэлээр ярьдаг	Mongolian
म नेपाली बोल्छु	Nepali
Mówię po polsku	Polish
Falo Portugues	Portuguese
ਮੈਂ ਪੰਜਾਬੀ ਬੋਲਦਾ ਹਾਂ	Punjabi
زه پښتو خبرې کولای شم	Pushto
Vorbesc limba română	Romanian
Я говорю по-русски	Russian
Ја говорим српски	Serbian
Ndino taura Shona	Shona
මම සිංහල භාෂාව කථාකරමි	Sinhalese
Rozprávam po slovensky	Slovak
Waxan ku hadlaa af Soomaali	Somali
Hablo español	Spanish
Ninasema Kiswahili	Swahili
Marunong ako magsalita ng Tagalog	Tagalog
நான் பேசும் மொழி தமிழ்	Tamil
ผมพูดไทย	Thai
నేను తెలుగు మాట్లాడతాను	Telugu
ትግርኛ እዛረብ እየ።	Tigrignia
Türkçe konuşuyorum	Turkish
Meka Twi	Twi
Я розмовляю по-українськи	Ukranian
میں اُردو بول سکتا ہوں	Urdu
Мен ўзбекча гапираман	Uzbeck
Chúng tôi nói tiếng Việt	Vietnamese
mo lè so yoruba	Yoruba

Ziel und Projektaufgaben

Ziele

- Ein System entwickeln, das die Sprache eines elektronischen Textdokuments identifizieren kann.

  Input (Dokument) → Output (Sprache)

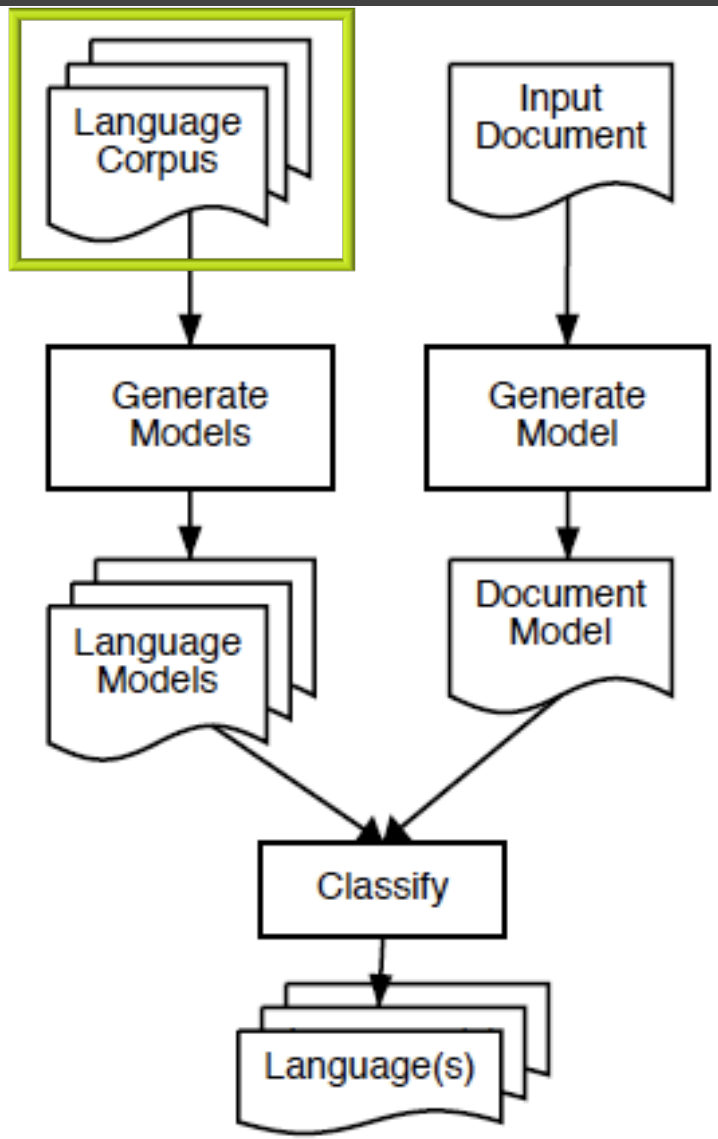
  - Wie groß muss das Trainings-/ Testdokument sein?
  - Welche Ansätze (N-Gramm/Token basiert, kombiniert) sind für welche Sprachen am besten geeignet?
  - Wie kann man eng verwandte Sprachen unterscheiden?
  - Welche spezielle Klassifizierer bzw. Klassifier-Kombinationen passen für schwierige Sprachabgrenzungen?

Aufgabe

- Trainings-/ Testdaten basierend auf Wikipedia
  - Tools zur Entfernung von Markups
- Implementierung der ausgewählten Ansätze
- Web-Interface erstellen
- Evaluierung / Dokumentation

Ablauf

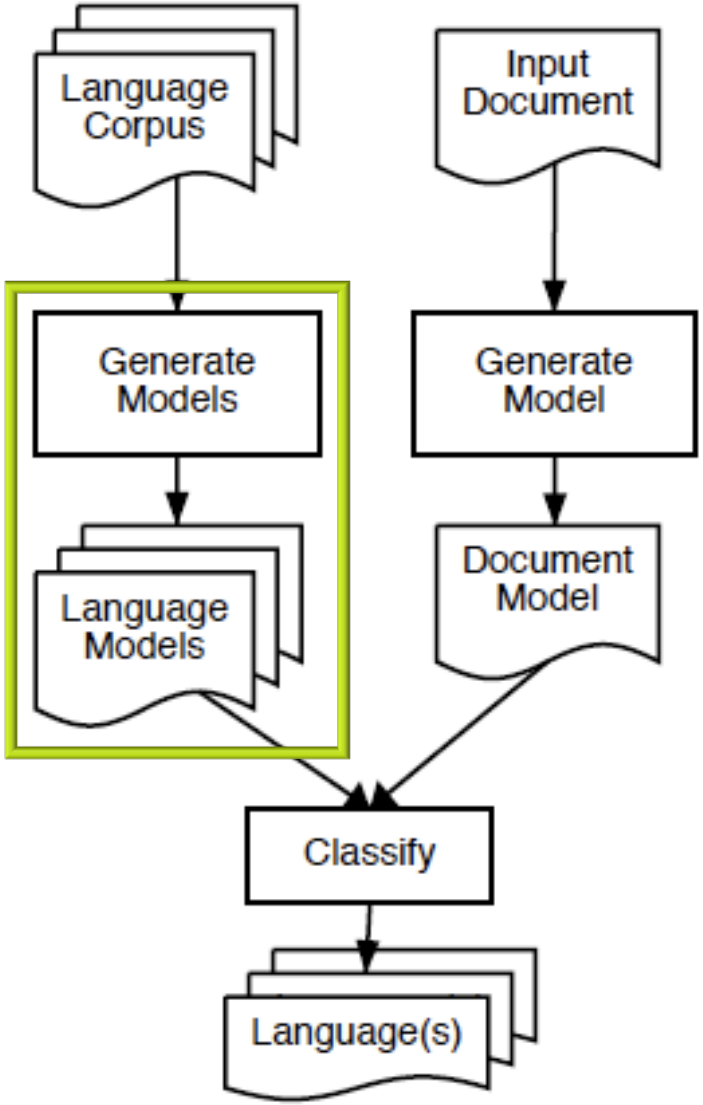
- Korpus aus Wikipediadaten
- Sprachmodelle erzeugen
- Testdokument
- Dokumentmodell erzeugen
- Dokumentmodell mit Sprachmodellen mittels verschiedener Klassifikationsverfahren vergleichen
- Wahrscheinlichste Sprache ausgeben



Quelle: Apoutsma 2001:2

Ablauf

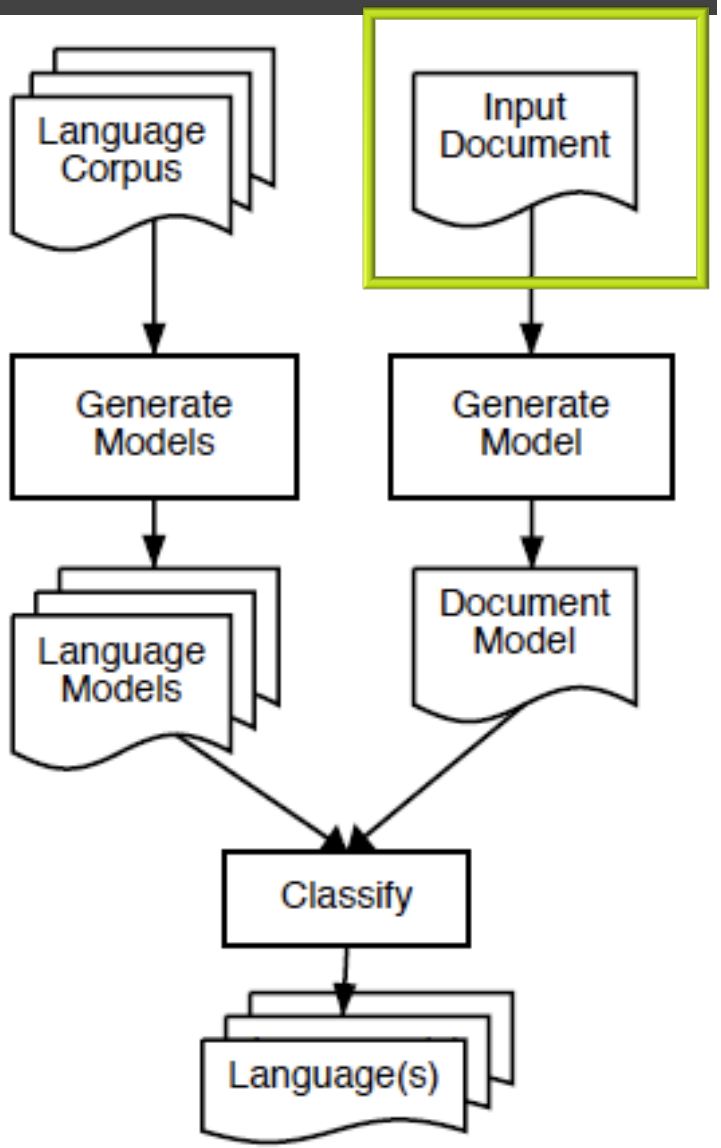
- Korpus aus Wikipediadaten
- Sprachmodelle erzeugen
- Testdokument
- Dokumentmodell erzeugen
- Dokumentmodell mit Sprachmodellen mittels verschiedener Klassifikationsverfahren vergleichen
- Wahrscheinlichste Sprache ausgeben



Quelle: Apoutsma 2001:2

Ablauf

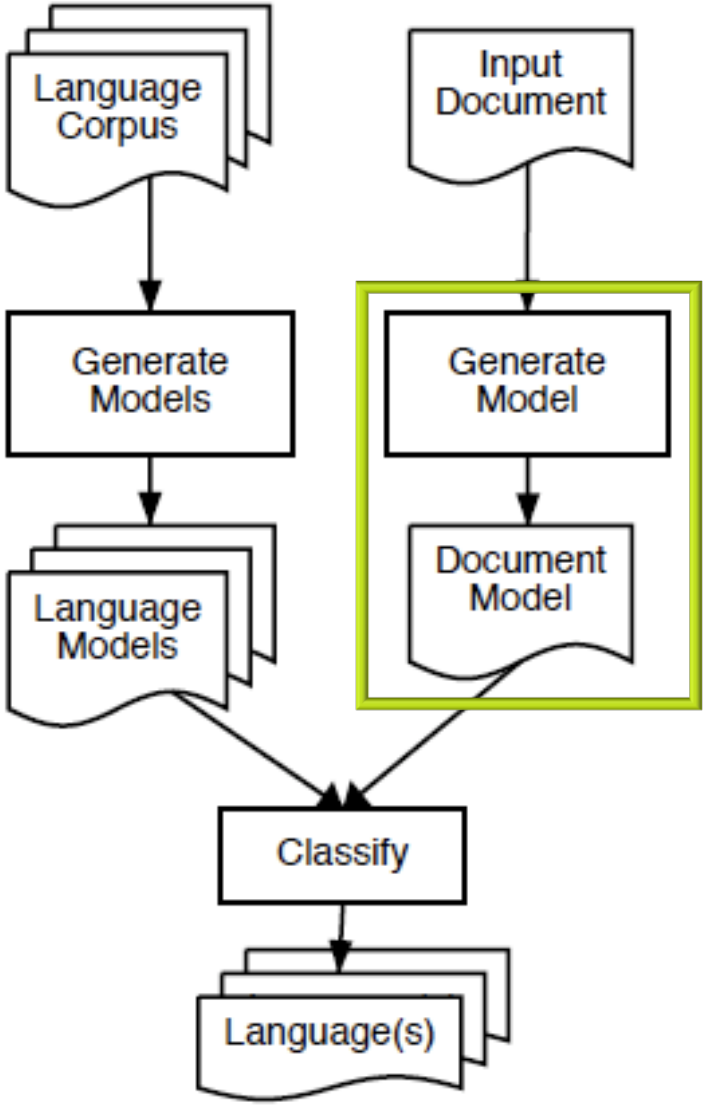
- Korpus aus Wikipediadaten
- Sprachmodelle erzeugen
- Testdokument
- Dokumentmodell erzeugen
- Dokumentmodell mit Sprachmodellen mittels verschiedener Klassifikationsverfahren vergleichen
- Wahrscheinlichste Sprache ausgeben



Quelle: Apoutsma 2001:2

Ablauf

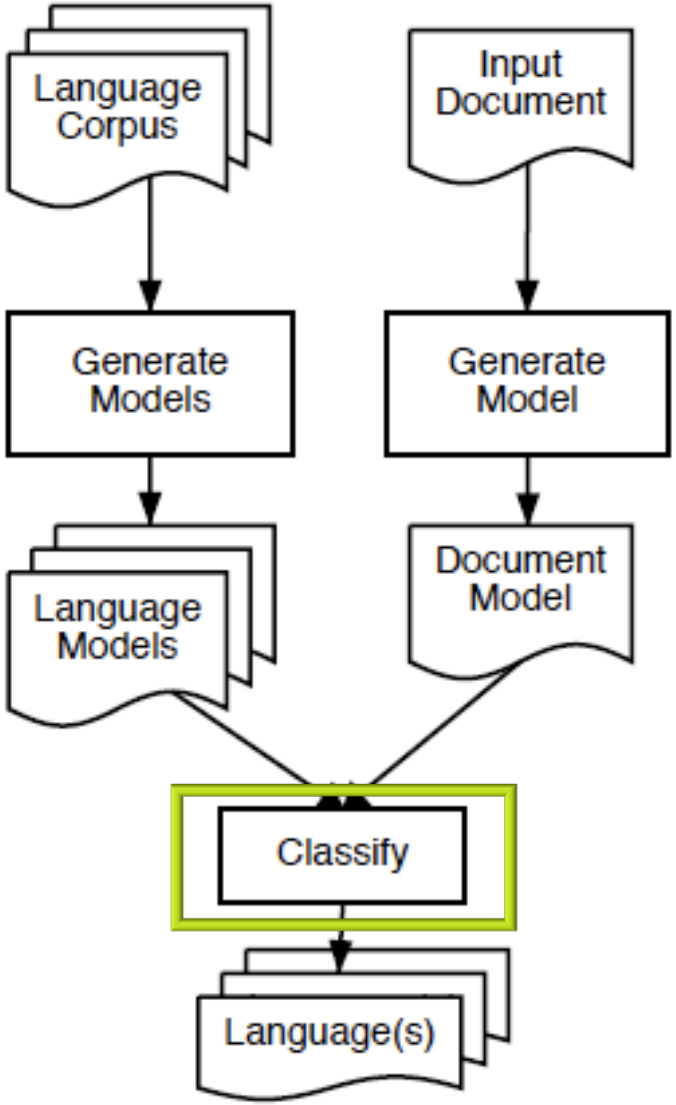
- Korpus aus Wikipediadaten
- Sprachmodelle erzeugen
- Testdokument
- Dokumentmodell erzeugen
- Dokumentmodell mit Sprachmodellen mittels verschiedener Klassifikationsverfahren vergleichen
- Wahrscheinlichste Sprache ausgeben



Quelle: Apoutsma 2001:2

Ablauf

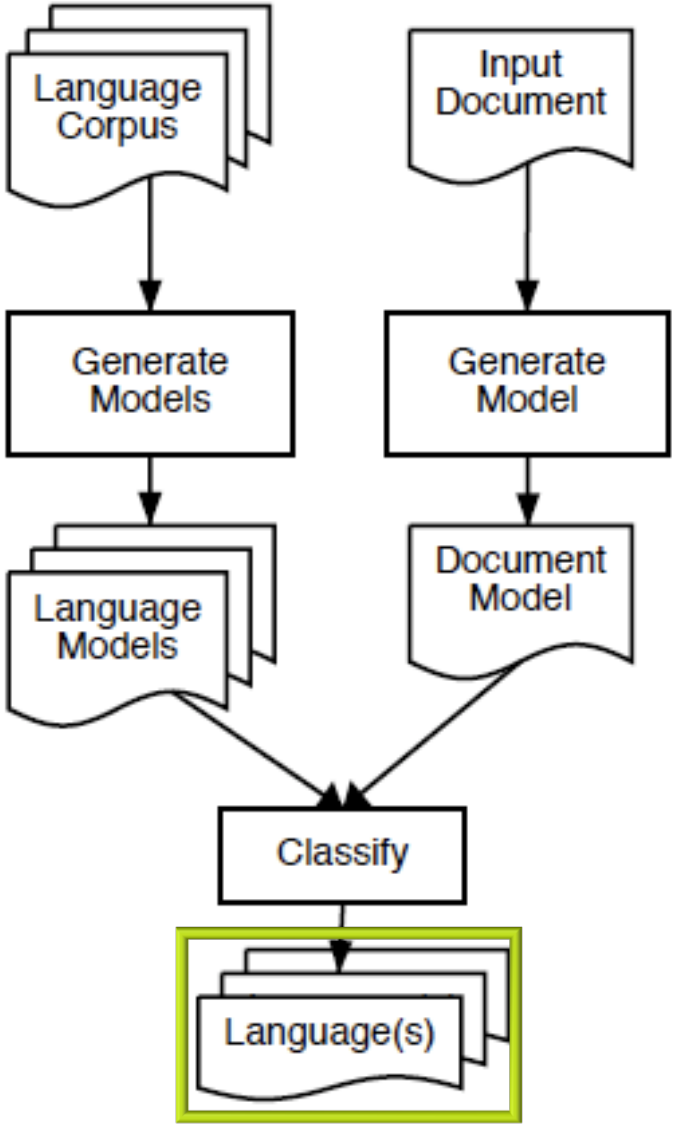
- Korpus aus Wikipediadaten
- Sprachmodelle erzeugen
- Testdokument
- Dokumentmodell erzeugen
- Dokumentmodell mit Sprachmodellen mittels verschiedener Klassifikationsverfahren vergleichen
- Wahrscheinlichste Sprache ausgeben



Quelle: Apoutsma 2001:2

Ablauf

- Korpus aus Wikipediadaten
- Sprachmodelle erzeugen
- Testdokument
- Dokumentmodell erzeugen
- Dokumentmodell mit Sprachmodellen mittels verschiedener Klassifikationsverfahren vergleichen
- Wahrscheinlichste Sprache ausgeben



Quelle: Apoutsma 2001:2

Stand der Forschung

- N-Gramme – William B. Canvar und John M. Trenkle (1994)
- Short words – J.M. Prager (1999), G. Grefnstette (1995)
- Frequent word – C. Souter (1994)
- Geschlossene Wortklassen R. D. Lins & P. Gonçalves (2004)
- Unikale Buchstabenkombinationen („sch“ im Deutschen) – C . Souter (1994)
- Word Shape Tokens – P .Sibun & L. A. Spitz (1994) und P. Sibun & J. C. Reynar(1996)

Ansätze

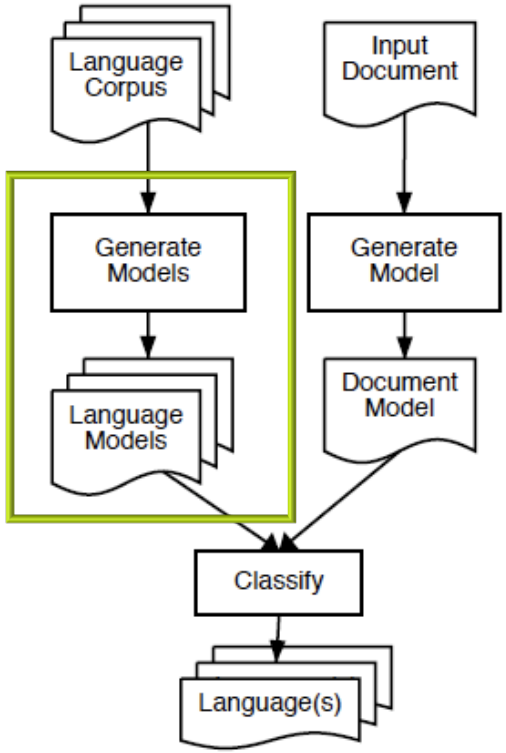
I. **Step By Step Ansatz :**
Language Identification With Confidence Limits (David Elworthy 1998)

II. **Character N-Gramm basierter Ansatz:**
(William B. Canvar und John M. Trenkle 1994)

III. **Kombinierter Ansatz:**
Linguini: Language Identification for Multilingual Documents (Prager, J. M. 1999)

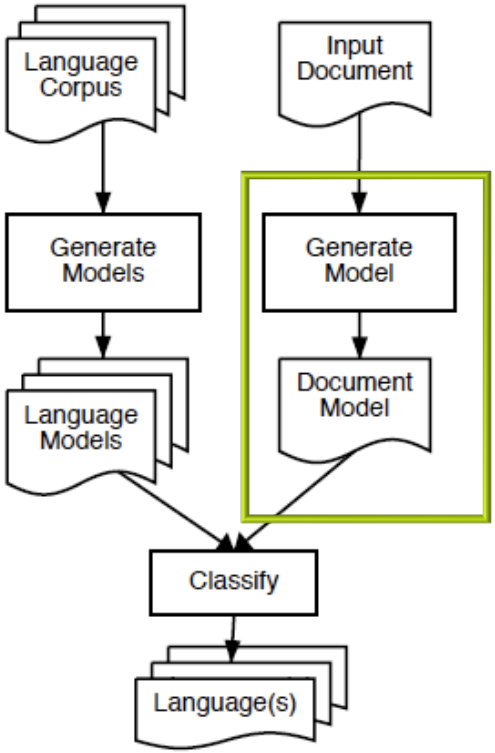
Step By Step Ansatz

- Tokenisieren der Trainingsdaten
- Berechnung der relativen Wahrscheinlichkeit des Tokens zu jedem Sprachmodell
- Tokenisieren des Testdokuments
- Erstellung des Testdokumentmodells nicht erforderlich
- Die wahrscheinlichste Sprache des Testdokuments mit Hilfe Bayes‘schen Entscheidungsregeln ermitteln



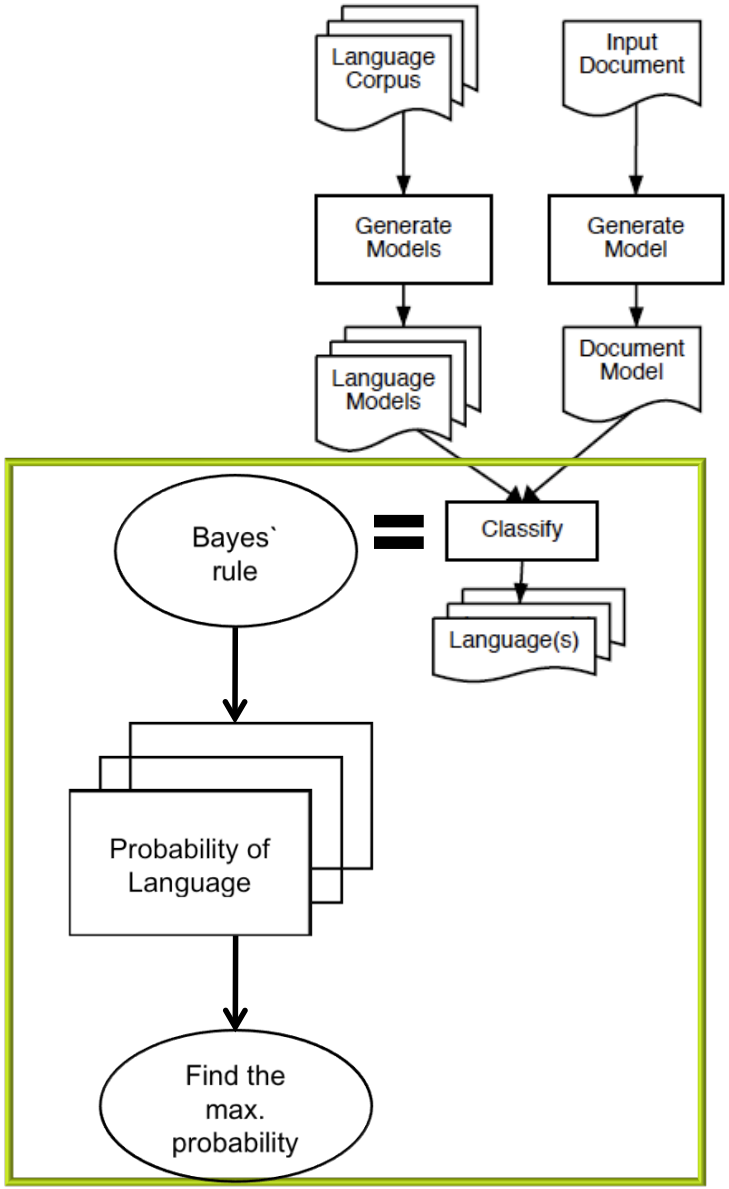
Step By Step Ansatz

- Tokenisieren der Trainingsdaten
- Berechnung der relativen Wahrscheinlichkeit des Tokens zu jedem Sprachmodell
- Tokenisieren des Testdokuments
- Erstellung des Testdokumentmodells nicht erforderlich
- Die wahrscheinlichste Sprache des Testdokuments mit Hilfe Bayes‘schen Entscheidungsregeln ermitteln



Step By Step Ansatz

- Tokenisieren der Trainingsdaten
- Berechnung der relativen Wahrscheinlichkeit des Tokens zu jedem Sprachmodell
- Tokenisieren des Testdokuments
- Erstellung des Testdokumentmodells nicht erforderlich
- Die wahrscheinlichste Sprache des Testdokuments mit Hilfe Bayes‘schen Entscheidungsregeln ermitteln



Step By Step Ansatz

- Klassifikation mittels Bayes'schen Entscheidungsregel

$$p(l|t) = \frac{p(t|l)p(l)}{p(t)}$$

Quelle: D.Elworthy. 1998: 2

- ***p(t|l)*** = relative Wahrscheinlichkeit eines Tokens zu jedem Sprachmodell
- ***p(l)*** = Wahrscheinlichkeit der Sprache
  - konstant
  - ***p(l)*** = $\frac{1}{\text{Anzahl der Sprachmodelle}}$
- ***p(t)*** = Wahrscheinlichkeit des Tokens über alle Sprachmodelle

Step By Step Ansatz

- Klassifikation mittels Bayes'schen Entscheidungsregel

$$p(l|t) = \frac{p(t|l)p(l)}{p(t)}$$

Quelle: D.Elworthy. 1998: 2

- Die Identifikation der Sprache passiert ohne dass das komplette Dokument angeschaut werden muss.
- Wenn genug Daten gesammelt wurden, um die Entscheidung zu treffen, bricht der Algorithmus ab.

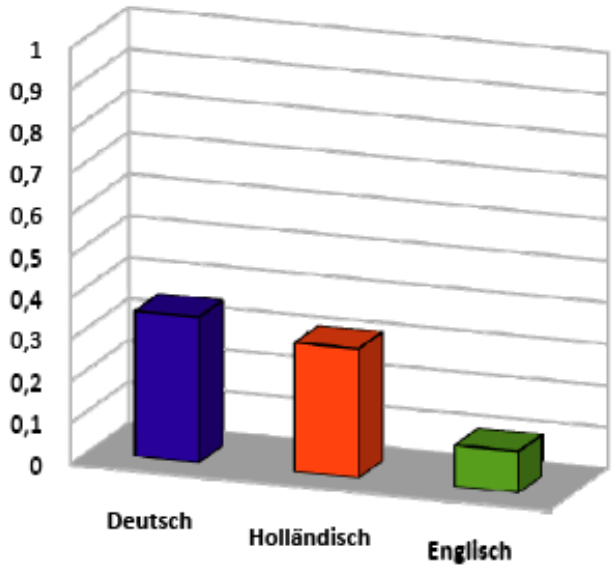
Step By Step Ansatz

- Beispiel

Das ist ein deutscher Satz.

Das

dt.: 0,35
holl.: 0,31
en.: 0,1



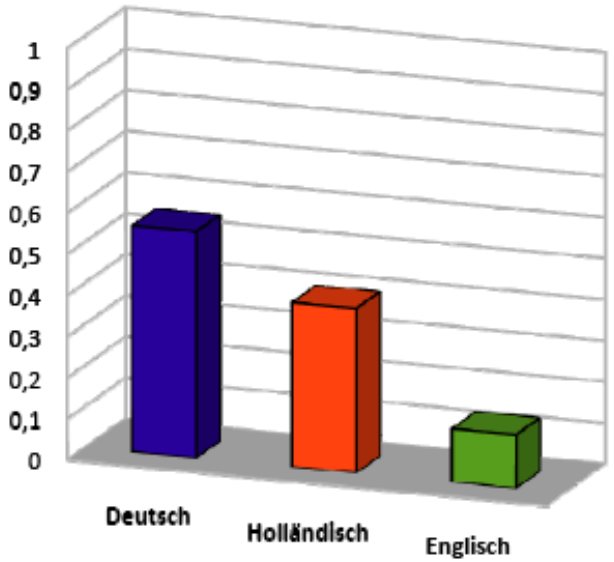
Step By Step Ansatz

- Beispiel

Das ist **ein deutscher Satz.**

dt.: 0,35
holl.: 0,31
en.: 0,1

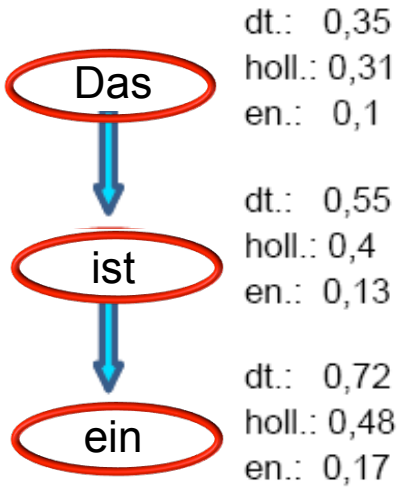
dt.: 0,55
holl.: 0,4
en.: 0,13

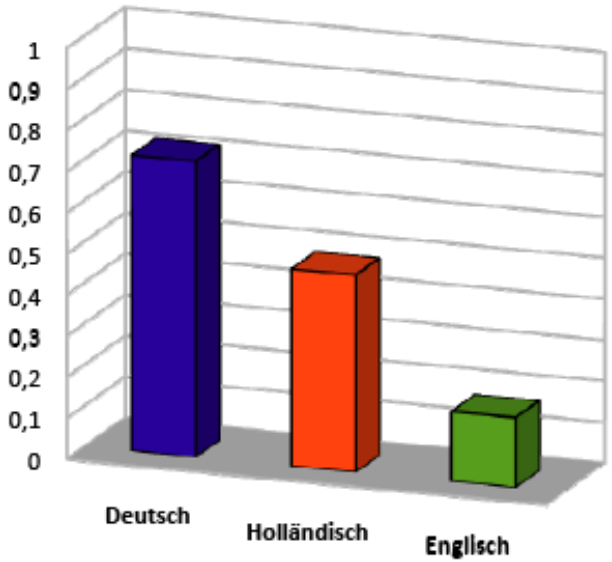


Step By Step Ansatz

- Beispiel

Das ist ein **deutscher Satz.**

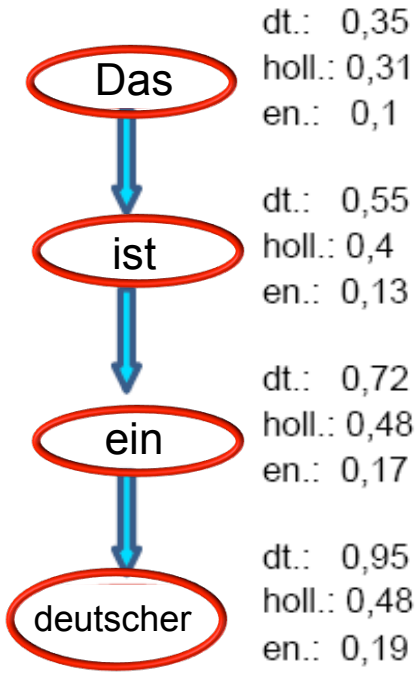




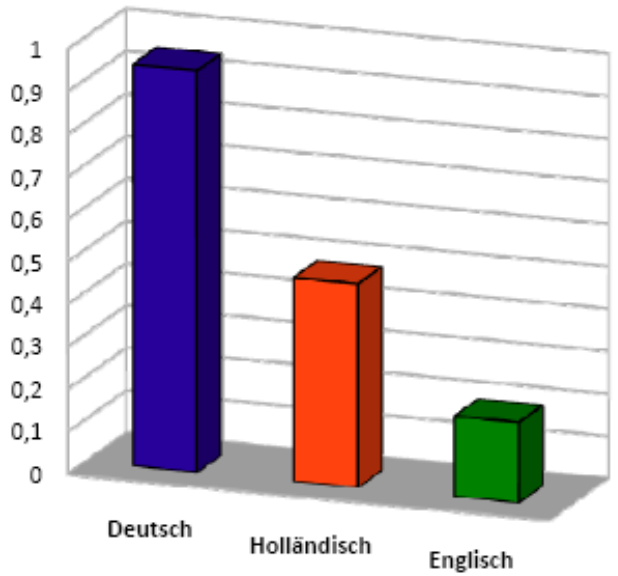
Step By Step Ansatz

- Beispiel

Das ist ein deutscher **Satz.**

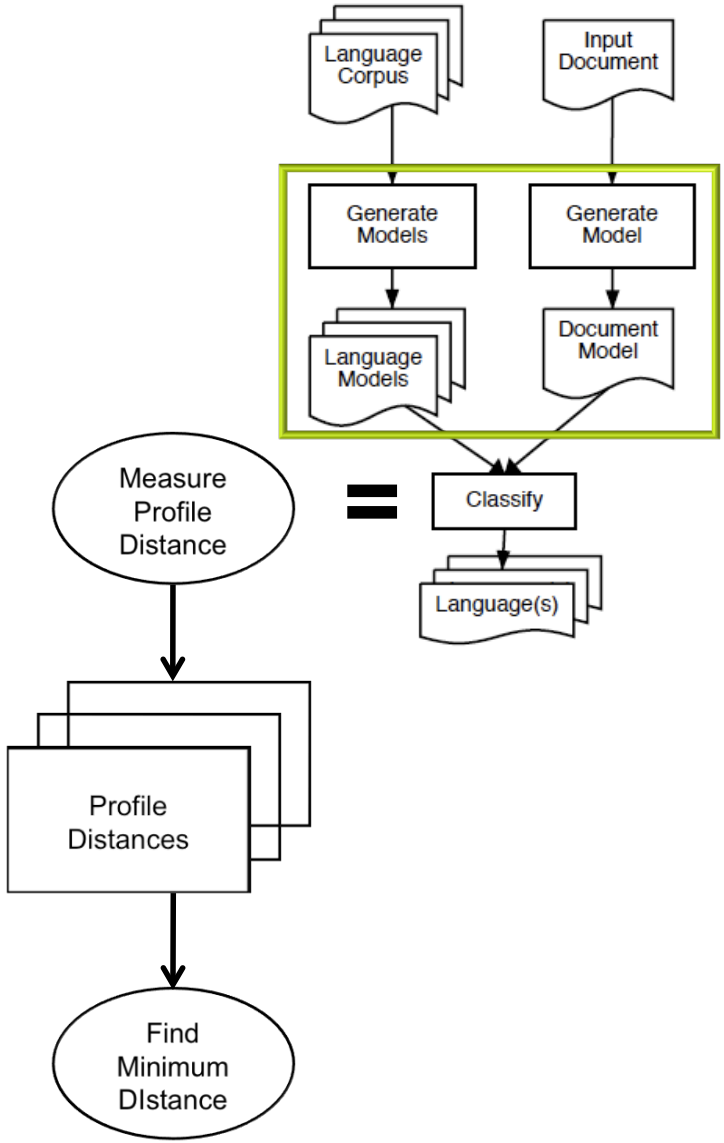


brake



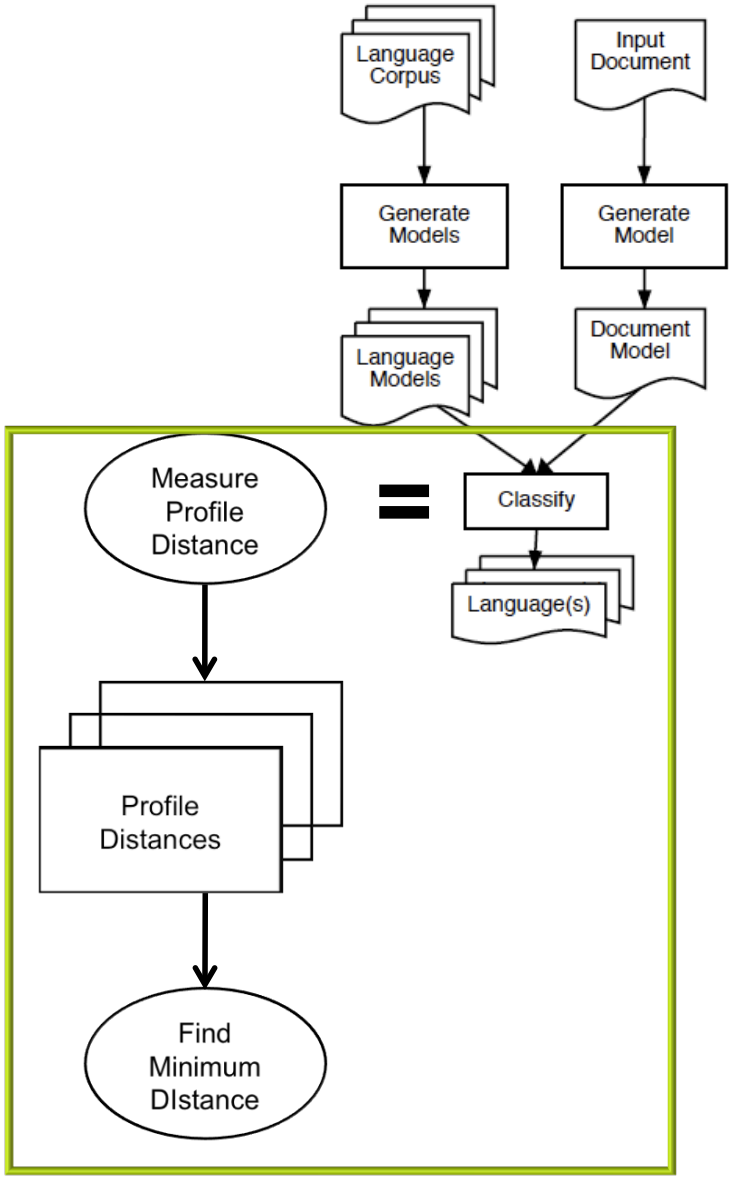
Character N-Gramm basierter Ansatz

- Tokenisieren, Zerlegen in N-Gramme, Berechnung der absoluten Häufigkeit der N-Gramme
- Sortierung der N-Gramme nach absoluten Häufigkeit absteigend
- Vergleich des Dokumentmodells mit dem Sprachmodell mittels Rangliste (Out-Of-Place measure)
- Die Sprache mit dem kleinsten Distanzwert ausgeben



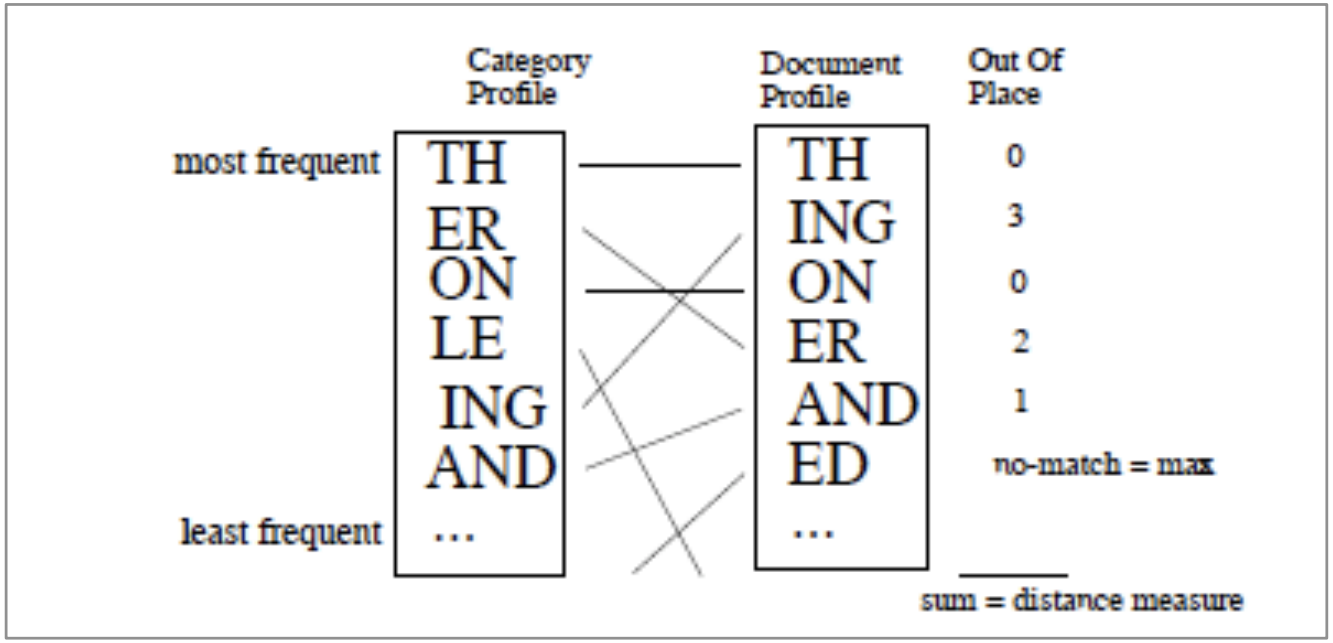
Character N-Gramm basierter Ansatz

- Tokenisieren, Zerlegen in N-Gramme, Berechnung der absoluten Häufigkeit der N-Gramme
- Sortierung der N-Gramme nach absoluten Häufigkeit absteigend
- Vergleich des Dokumentmodells mit dem Sprachmodell mittels Rangliste (Out-Of-Place measure)
- Die Sprache mit dem kleinsten Distanzwert ausgeben



Character N-Gramm basierter Ansatz

Ad hoc Ranking – Out-Of-Place measure:



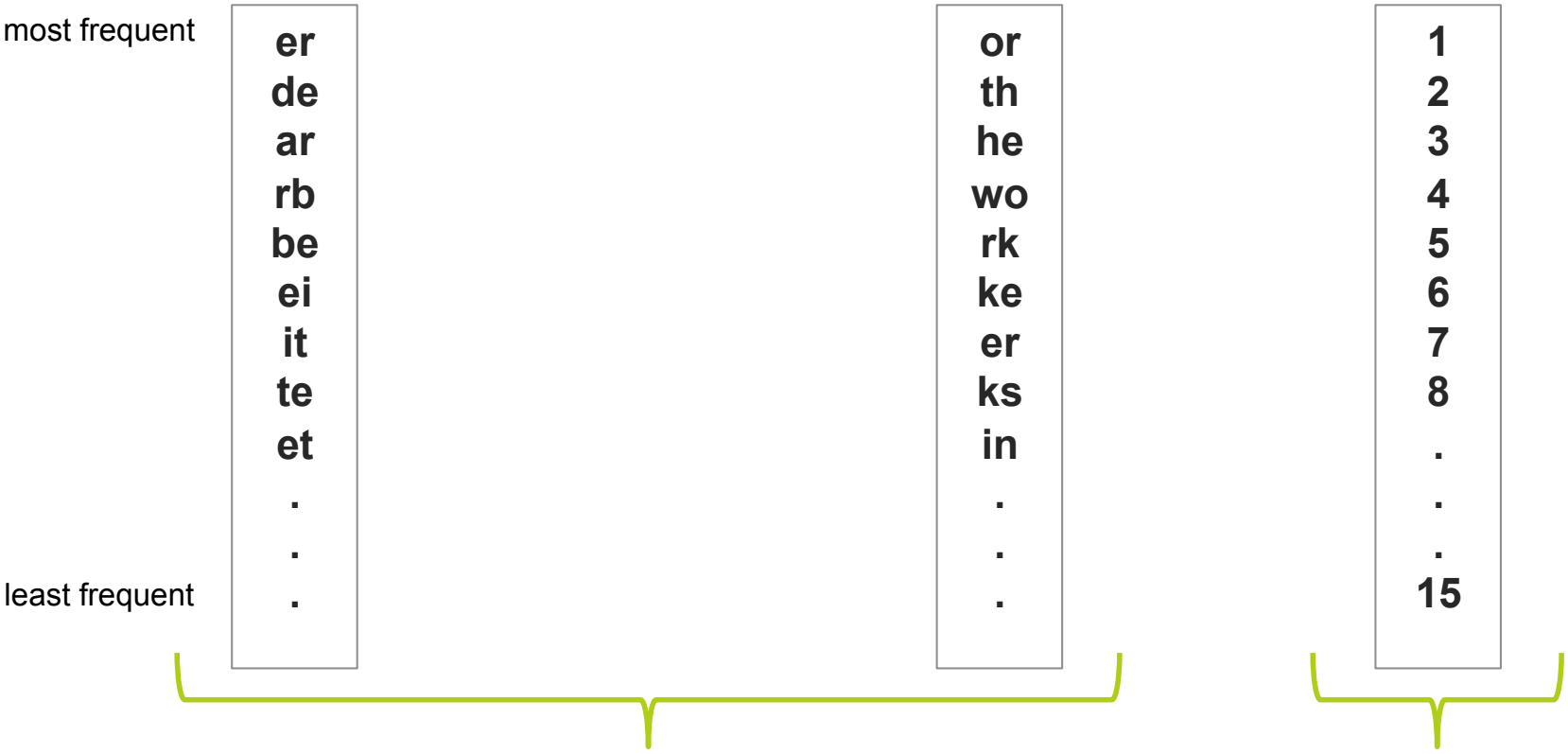
Quelle: B. Cavnar and M. Trenkle 1994:6

Character N-Gramm basierter Ansatz

- Beispiel

Der Arbeiter arbeitet in der Fabrik

The worker works in the factory



Character N-Gramm basierter Ansatz

- Beispiel

Sprachmodell (Deutsch):
Der Arbeiter arbeitet in der Fabrik

Testdokument:
Er leitet die Worte weiter

Rank	Sprachmodell	Dokumentmodell	Out-Of-Place
1	er	te	**7**
2	de	ei	**4**
3	ar	er	**2**
4	rb	it	**3**
5	be	di	**max = 15**
6	ei	et	**3**
7	it	ie	**max**
8	te	le	**max**
9	.	.	**.**
10	.	.	**.**
11	.	.	**.**
12	.	.	**.**
13	.	.	**.**
14	.	.	**.**
15	.	.	**.**
			Summe =123

Character N-Gramm basierter Ansatz

- Beispiel

Sprachmodell (Englisch):
The worker works in the factory

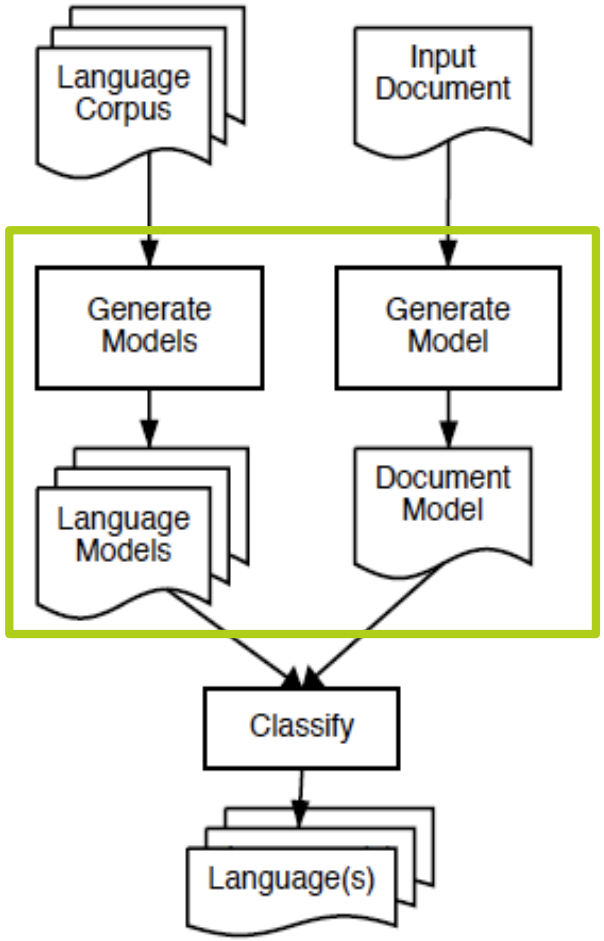
Testdokument:
Er leitet die Worte weiter

Rank	Sprachmodell	Dokumentmodell	Out-Of-Place
1	or	te	**max = 15**
2	th	ei	**max**
3	wo	er	**3**
4	rk	it	**max**
5	ke	di	**max**
6	er	et	**max**
7	ks	ie	**max**
8	in	le	**max**
9	.	.	**.**
10	.	.	**.**
11	.	.	**.**
12	.	.	**.**
13	.	.	**.**
14	.	.	**.**
15	.	.	**.**
			Summe =196

Kombinierter Ansatz

N-Gramm-Wort-kombinierter Ansatz

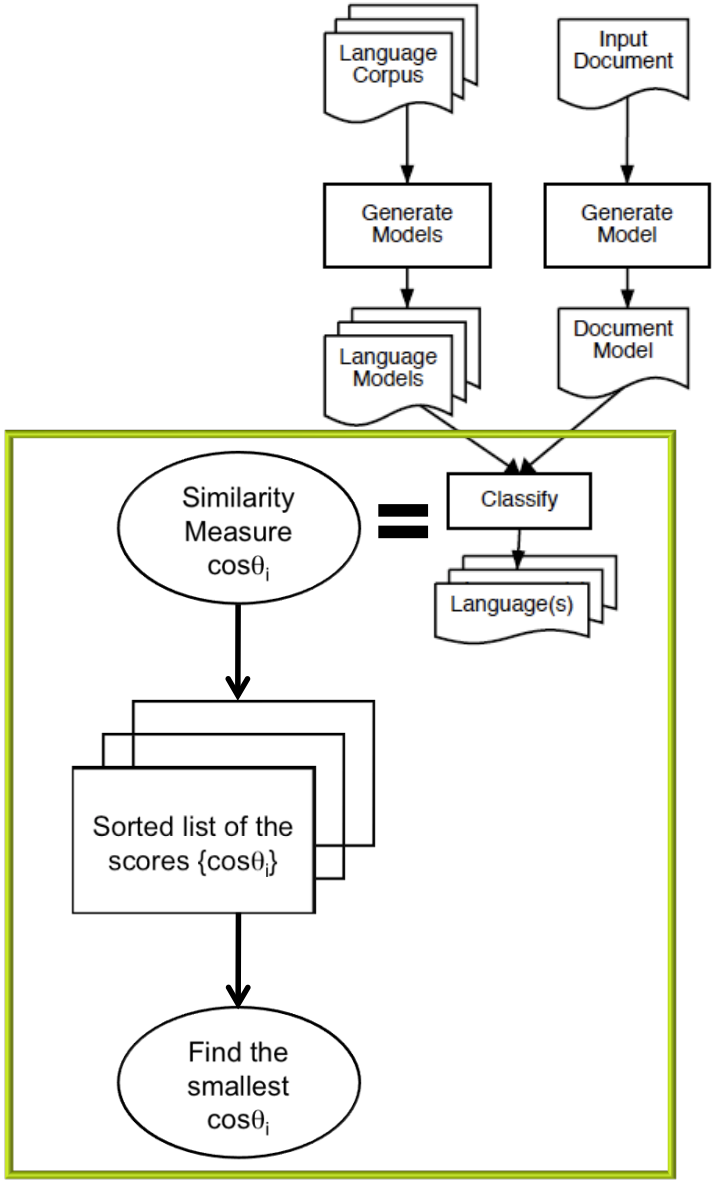
- Zerlegen:
  - in N-Gramme (n= 2:5) : Bi-Gramme, Tri-Gramme, Quad-Gramme, Pento-Gramme
  - in Tokens
  - in kurze Wörter (w = 1:4) + Tri-Gramme
  - in kurze Wörter (w = 1:4) + Quad-Gramme
  - in Wörter der unbegrenzten Länge + Quad-Gramme
- Generierung der Dictionaries



Kombinierter Ansatz

Vektorraummodell

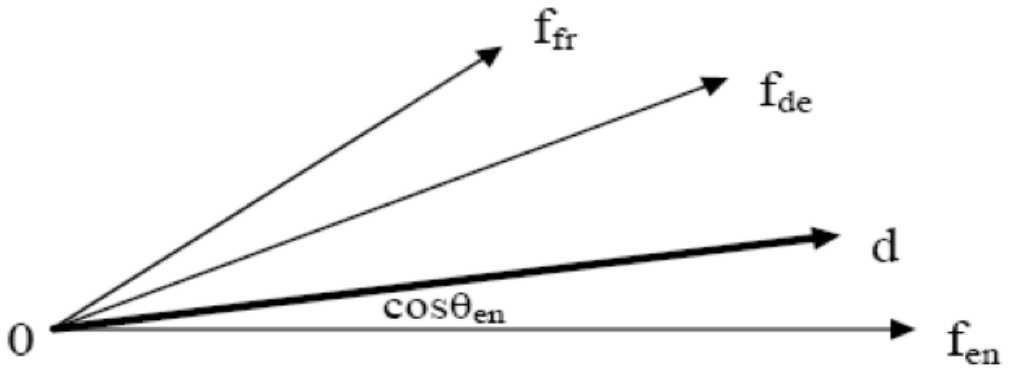
- Darstellung des Dokumentmodells durch Vektor $\vec{d}$ und des Sprachmodells durch Vektor $\vec{f}_j$
- Ermittlung der Ähnlichkeit zwischen dem Dokumentmodellvektor und den Sprachmodellvektoren mittels der Kosinusdistanz
- Ausgabe der geordneten Liste von Sprachen (hit-list) mit Prozentangaben



Kombinierter Ansatz

- Berechnung des Ähnlichkeitsmaßes eines Vektors in einem mehrdimensionalen Raum

$$\cos\theta_i = \frac{\vec{d} \cdot \vec{f}_i}{\left|\vec{d}\right| \cdot \left|\vec{f}_i\right|}$$

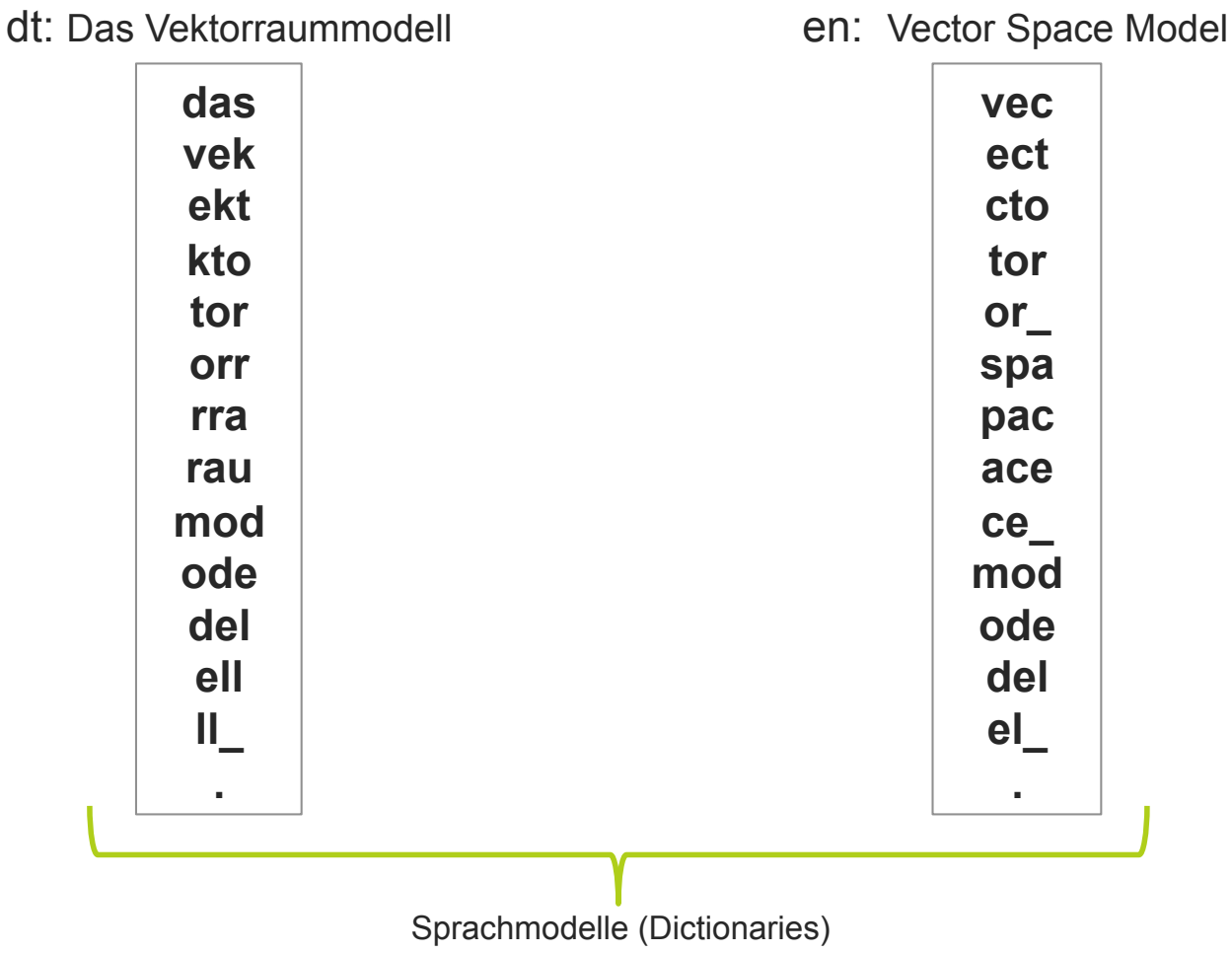


d – Dokumetmodellvektor
f – Sprachmodellvektor

Quelle: Prager J. M. 199: 7

Kombinierter Ansatz

- Beispiel



Kombinierter Ansatz

- Vektoren für das Dokumentmodell

(*das, tor, ect, mod, ode, del, ell*):

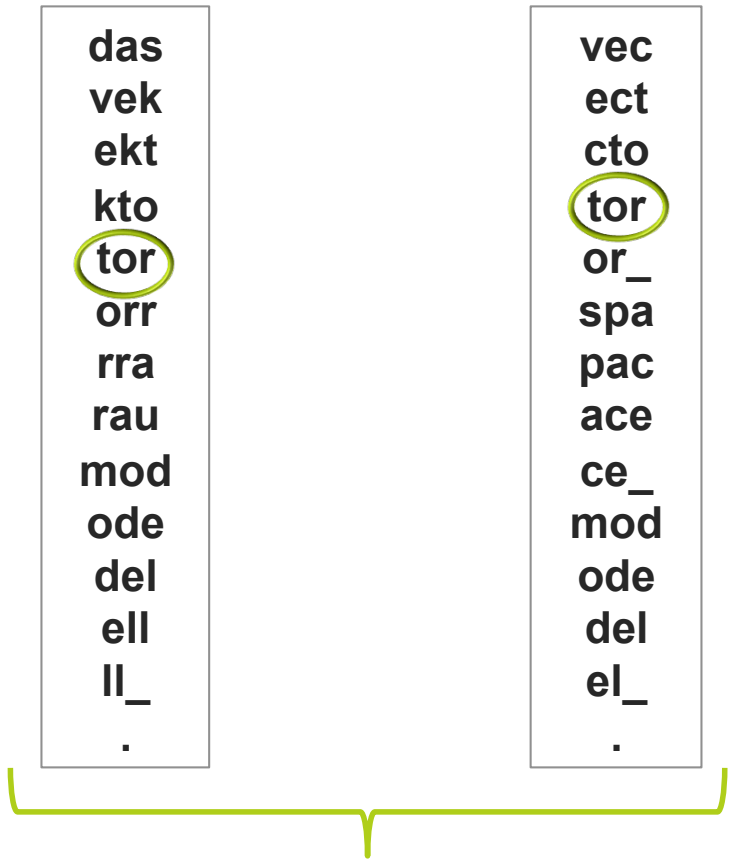
- Merkmalsvektor: (1,1,1,1,1,1,1)
- Englisch: (0,1,1,1,1,1,0)

$$cos\theta_i = \frac{5}{\sqrt{35}} = 0.84$$

- Deutsch: (1,1,0,1,1,1,1)

$$cos\theta_i = \frac{6}{\sqrt{42}} = 0.92$$

dt: Das Vektorraummodell en:Vector Space Model



Ressourcen

- Es gibt in Wikipedia 10 Millionen Artikel in 250 Sprachen.

100 000+

العربية • Català • Česky • Dansk • Deutsch • English • Español • Esperanto • Français • Bahasa Indonesia • Italiano • Magyar • Nederlands • 日本語 • 한국어 • Norsk (bokmål) • Polski • Português • Русский • Română • Slovenčina • Suomi • Svenska • Türkçe • Українська • Tiếng Việt • Volapük • 中文

10 000+

Afrikaans • Aragonés • Asturianu • Kreyòl Ayisyen • Azərbaycan / آذربايجان ديلی • □□□□□ • Беларуская (Акадэмічная ·Тарашкевіца) • □□□□□□□□□□□□□ □□□□□□□ • Bosanski • Brezhoneg • Български • Чăваш • Cymraeg • Eesti • Ελληνικά • Euskara • فارسی • Frysk • Galego • ગુજરાતી • हिन्दी • Hrvatski • Ido • Íslenska • עברית • Basa Jawa • ქართული • Kurdî / كوردی • Latina • Latviešu • Lëtzebuergesch • Lietuvių • Македонски • □□□□□□ • मराठी • Bahasa Melayu • नेपाल भाषा • Norsk (nynorsk) • Nnapulitano • Occitan • Piemontèis • Plattdüütsch • Ripoarisch • Runa Simi • Shqip • Sicilianu • Simple English • Sinugboanon • Slovenščina • Српски • Srpskohrvatski / Српскохрватски • Basa Sunda • Kiswahili • Tagalog • தமிழ் • □□□□□□ • ไทย • اردو • Walon • Winaray • 粵語 • Žemaitėška

1 000+

Alemannisch • □□□□ • Armãneashce • Arpitan • □□□□□□ • Avañe'ẽ • Bân-lâm-gú • Basa Banyumasan • भोजपुरी • Bikol Central • Boarisch • བོད་ཡིག • Corsu • Deitsch • □□□□□□ • Eald Englisc • Эрзянь • Fiji Hindi • Føroyskt • Furlan • Gaeilge • Gaelg • Gàidhlig • 贛語 • گیلکی • 文言 • Hak-kâ-fa / 客家話 • ʻŌlelo Hawaiʻi • Հայերեն • Hornjoserbsce • Ilokano • Interlingua • Interlingue • Ирон Æвзаг • □□□□□ • Kapampangan • Kaszëbsczi • Kernewek • □□□□□□□□ • Коми • Ladino / לאדינו • Ligure • Limburgs • Lingála • lojban • Lumbaart • Malagasy • Malti • Māori • مصرى • مازِرونی / Mäzeruni • Монгол • □□□□ • Nāhuatlahtōlli • Nedersaksisch • नेपाली • Nouormand • Novial • O'zbek • पाऴि • Pangasinán • پښتو • ਪੰਜਾਬੀ / پنجابی • Қазақша • Qırımtatarca • Rumantsch • संस्कृतम् • Sámegiella • Sardu • Саха Тыла • Scots • Seeltersk • شاہ مکھی پنجابی • □□□□□ • Ślůnski • Tarandíne • татарча / Tatarça • Тоҷикӣ • Lea faka-Tonga • Türkmen • Uyghur / ئۇيغۇرچە • Vèneto • Võro • West-Vlams • Wolof • 吴语 • ייִדיש • Yorùbá • Zazaki

100+

Bahsa Acèh • Аҧсуа • Авар • Aymara • Bamanankan • Башҡорт • Bislama • Буряад • Chamoru • Chavacano de Zamboanga • Cuengh • Diné Bizaad • Dolnoserbski • Emigliàn-Rumagnòl • Estremeñu • Eʋegbe • □□□□□□ • Хальмг • Hausa / هَوُسَا • Igbo • ᐃᓄᒃᑎᑐᑦ / Inuktitut • Kalaallisut • कश्मीरी / كشميري • Kongo • Кыргызча • □□□□□□□ • Mìng-dĕ̤ng-ngṳ̄ • Mirandés • Мокшень • Молдовеняскэ • Dorerin Naoero • Nēhiyawēwin / ᓀᐦᐃᔭᐍᐏᐣ • Norfuk / Pitkern • Нохчийн • Олык Марий • □□□□□ • Afaan Oromoo • □□□□□□□ • Papiamentu • Ποντιακά • Qaraqalpaqsha • ཇོང་ཁ • Romani / रोमानी • Gagana Sāmoa • Setswana • سنڌي • Словѣньскъ / □□□□□□□□□□ • Af Soomaali • کوردی • SiSwati • Sranantongo • Reo Tahiti • Taqbaylit • Tetun • □□□□ • Tok Pisin • ᏣᎳᎩ • □□ □□□□ • Xitsonga • Удмурт • Tshivenḓa • isiXhosa • Zeêuws • isiZulu

Other languages • Weitere Sprachen • Autres langues • Kompletna lista języków • 他の言語 • Otros idiomas • 其他语言 • Другие языки • Aliaj lingvoj • 다른 언어 • Ngôn ngữ khác

Evaluation

- Eigenschaften von Dokumenten und Sprachen herausfinden;
- Trainings-/Testdaten unterschiedlicher Länge verwenden;
- Eine zu starke Anpassung an die Trainingsdaten vermeiden;
- Die Kodierung für Dokumente in nicht lateinischen oder mit einem erweiterten lateinischen Zeichensätzen korrekt erkennen;
- N-Gramm Längen n=1:5/ kurze Wörter/ Wörter unterschiedlicher Längen verwenden;
- Den Ansatz für die Identifikation der eng verwandten Sprachen bestimmen;
- Bei schwierigen Sprachabgrenzungen versuchen Klassifizierer zu kombinieren.

Aufgabenverteilung und Zeitplan

November		December		Januar		Februar	
01.11.09	Thema einarbeiten & Spezifikationsvortrag vorbereiten[Alle]	01.12.09	Implementierung- 1 Ansatz[Galina] 2 Ansatz[Ulzhan] 3 Ansatz [George/Maria]	01.01.10	Training & testen [alle]	01.02.10	
02.11.09		02.12.09		02.01.10		**02.02.10**	**Kurzes Demo**
03.11.09		03.12.09		03.01.10		03.02.10	Dokumentation & Webinterface [Galina/Maria]
04.11.09		04.12.09		04.01.10		04.02.10	
05.11.09		05.12.09		05.01.10		05.02.10	
06.11.09		06.12.09		06.01.10		06.02.10	
07.11.09		07.12.09		07.01.10		07.02.10	
08.11.09		08.12.09		08.01.10	Evaluation & Fehleranalyse & Verbesserung [alle]	08.02.10	
09.11.09		09.12.09		09.01.10		09.02.10	
10.11.09		10.12.09		10.01.10		10.02.10	
11.11.09		11.12.09		11.01.10		11.02.10	
12.11.09		12.12.09		12.01.10		12.02.10	
13.11.09		13.12.09		13.01.10		13.02.10	
14.11.09		14.12.09		14.01.10		14.02.10	
15.11.09		15.12.09		15.01.10		15.02.10	
16.11.09		16.12.09		16.01.10		16.02.10	
17.11.09		17.12.09		17.01.10		17.02.10	
18.11.09		18.12.09		18.01.10		18.02.10	
19.11.09		19.12.09		19.01.10		**19.02.10**	**Systemabgabe**
20.11.09		20.12.09		20.01.10		20.02.10	Poster erstellen [George/Ulzhan]
21.11.09		21.12.09		21.01.10		21.02.10	
22.11.09		22.12.09	Training & testen [alle]	22.01.10		22.02.10	
23.11.09		23.12.09		23.01.10		23.02.10	
24.11.09	**Spezifikationsvortrag**	24.12.09		24.01.10		24.02.10	
25.11.09	Korpus vorbereiten [Alle]	25.12.09		25.01.10		25.02.10	
26.11.09		26.12.09		26.01.10		26.02.10	
27.11.09		27.12.09		27.01.10		27.02.10	
28.11.09		28.12.09		28.01.10		28.02.10	
29.11.09		29.12.09		29.01.10			
30.11.09		30.12.09		30.01.10			
		31.12.09		31.01.10			

Quellen

- W. Cavnar and J.M. Trenkle. 1994. N-gram-based text categorization.
- T. Dunning. 1994. Statistical identification of language.
- D. Elworthy. 1998. Language identification with confidence limits.
- G. Grefenstette. 1995. Comparing two language identificationschemes.
- J. Prager 1999. Linguini: Language Identification for Multilingual Documents
- P. Sibun and A.L. Spitz. 1994. Language determination: Natural language processing from scanned document images.

Danke für Eure Aufmerksamkeit