Homework 4

Fabian Otto – Matrikelnummer: 2792549
Foundations of Language Technology

19. November 2017

1 Erklärung des Algorithmus (4.1 b)

Der Algorithmus ist größtenteils im Code bereits kommentiert.
Für die Entwicklung eines Scores sollte die Frequenz der Character herangezogen werden. Entsprechend dem, ist es naheliegend eine übliches Distanzmaß heranzuziehen. Hierfür wurden im Rahmen dieser Hausübung die Euklidische Distanz, die Minkowski L1 Distanz und die Kosinus Ähnlichkeit herangezogen.
Euklidische Distanz:

$$\mathrm{d}(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

Minkowski L1 Distanz:

$$\mathrm{d}(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{n} |(p_i - q_i)|$$

Kosinus Ähnlichkeit für Vektoren x und y:

$$cos(\theta) = cos(\boldsymbol{x}, \boldsymbol{y}) = \frac{\sqrt{\sum_{i=1}^{n} x_i y_i}}{\sqrt{\sum_{i=1}^{n} x_i^2} \sqrt{\sum_{i=1}^{n} y_i^2}}$$

Es ergaben sich folgende Vorhersagen für die drei Distanzmaße:

Text	**Euklidische Distanz**	**Minkowski L1 Distanz**	**Kosinus Ähnlichkeit**
Englisch	Deutsch	English	Französisch
Französisch	Französisch	Französisch	Deutsch
Deutsch	Französisch	Deutsch	Englisch

Tabelle 1: Vorhersagen der verschiedenen Distanzmaße

Hier zeigt sich, dass die Verwendung der Minkowski L1 Distanz zu den besten Ergebnissen führt und wurde deshalb im Algorithmus als Scoring verwendet. Je geringer die Distanz desto näher ist die Input Text Sprache an der gegebenen Sprache.

Von der Grundidee iteriert der Algorithmus über alle Elemente in der *language_ base* und vergleicht die Frequenz der Character mit der Frequenz im vorliegenden Text. Dabei wird für jedes Element die Minkowski Distanz berechnet und aufsummiert. Allerdings muss hierbei berücksichtigt werden, dass im Text Elemente enthalten sein können, die nicht in der *language_ base* vorkommen, z.B. die englische *language_ base* beinhaltet kein *ö*, was in einem deutschen Text jedoch vorkommen kann. Somit müssen die Frequenzen von solchen Charactern am Ende noch hinzugefügt werden, da hierüber eine Abgrenzung zu anderen Sprachen stattfindet, wenn z.B. viele französische Sonderzeichen im Text enthalten sind un mit einer englischen *language_ base* verglichen werden, wird die Differenz erhöht und die Vorhersage von Englisch (für einen französischen Text) unwahrscheinlicher. Im Algorithmus wird die Suche nach solchen Chars mit der *difference* Methode auf die Sets von *language_ base* und FreqDist(input_text) ausgeführt.

2 Unterscheidung von englischen und deutschen Texten (4.1 e)

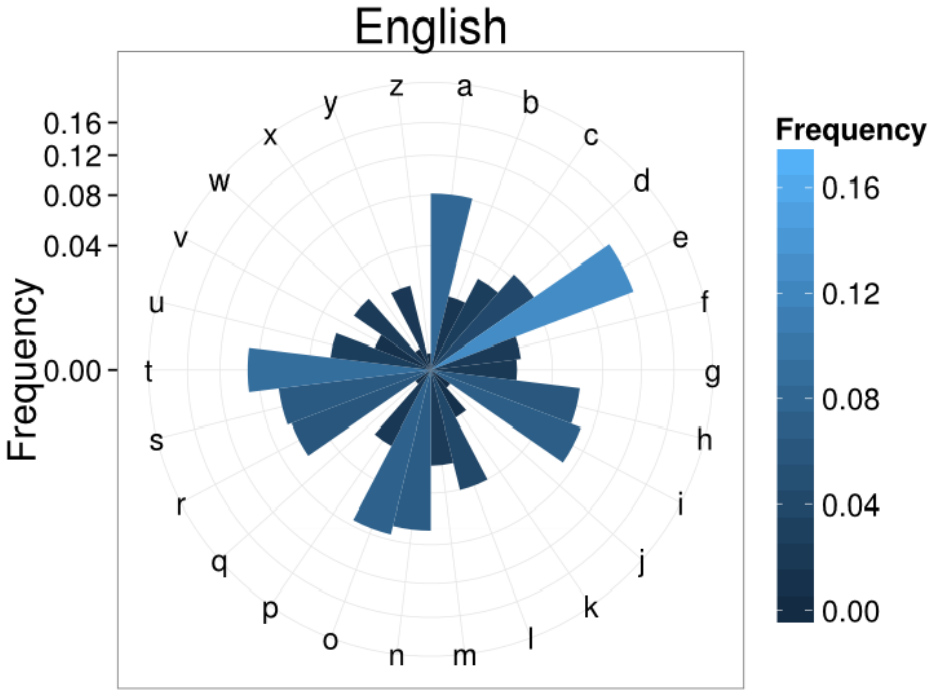

Abbildung 1: Frequenz der englischen Buchstaben, Quelle:`http://i.imgur.com/GE7LJR6.png`, aufgerufen am 19.11.2017

Wie aus den beiden Abbildungen 1 und 2 ersichtlich wird, ist die reine Betrachtung von Buchstaben nicht sinnvoll. Vor allem Buchstaben, wie *e*, *n*, *s*, *r* und *i* zeigen sowohl im Englischen als auch im Deutschen sehr große Ähnlichkeiten hinsichtlich der Häufigkeitsverteilungen auf. Damit is es nur schwerlich möglich eine exakte Bestimmung vorzunehmen, vor allem unter Berücksichtigung der kurzen Inputtexte, die wenig Aufschluss über die Spra-

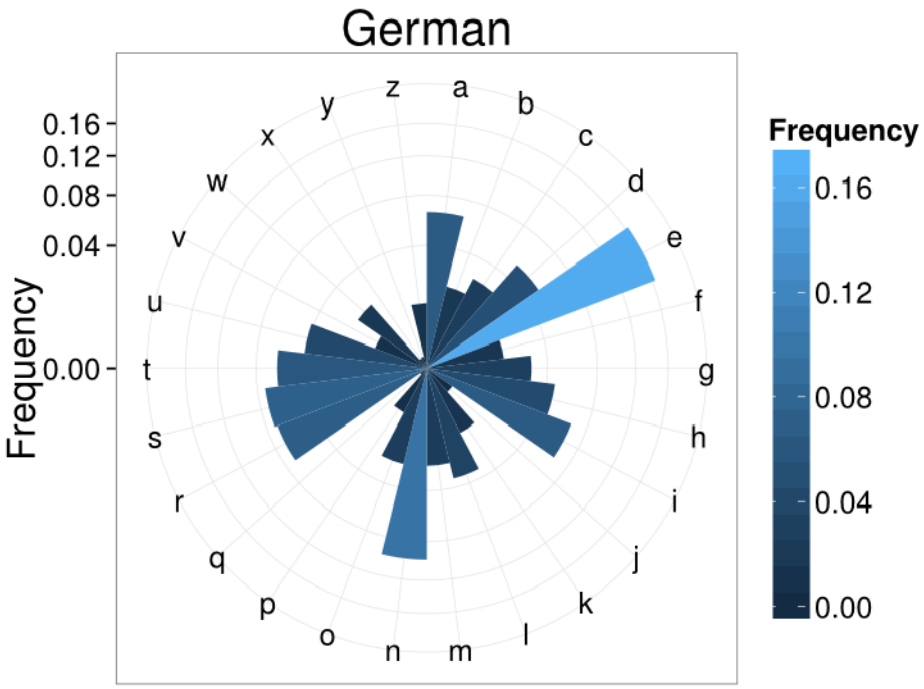

Abbildung 2: Frequenz der deutschen Buchstaben, Quelle:`http://i.imgur.com/GE7LJR6.png`, aufgerufen am 19.11.2017

che geben. Weiterhin ist die Anzahl an Sonderzeichen (namentlich ä,ö,ü,ß) im Deutschen sehr gering, wenn dies im Vergleich zu z.B. dem Französischen gesehen wird. Das heißt die Anzahl der Buchstaben unterscheidet sich zwischen Deutsch und English nur unmerklich. Damit wird aufgrund der Sonderzeichen keine größere Differenz im Scoring hervorgerufen.

3 Verwendung von Tokens, Character- und Tokenbigrammen (4.2 d)

Aus einer theoretischen Sicht macht die Verwendung von Tokenbigrammen wenig Sinn, da dabei angenommen werden müsste, dass die Texte ähnliche Inhalte und einen ähnlichen Stil verfolgen. Dies heißt die Wahrscheinlichkeit, dass ein Tokenbigramm in verschiedenen Texten überhaupt zweimal vorkommt, ist gering, noch geringer ist jedoch die Wahrscheinlichkeit, dass dieses mit der gleichen relativen Häufigkeit im gesamten Text zu finden ist. Einer Ausnahme stellen dabei 2-elementige Kollokationen dar, allerdings dürfte die Gesamtanzahl der Bigramme wohl keine Kollokationen beinhalten. Dies bestätigen auch die praktischen Ergebnisse. Weder mit der Euklidischen Distanz noch der Minkowski L1 Distantz noch der Kosinus Ähnlichkeit kann hiermit eine korrekte Vorhersage getroffen werden.

Die Verwendung von Tokens bringt dasselbe Problem mit sich, hierbei ist die Wahrscheinlichkeit zwar etwas höher, dass die Token erneut vorkommen. In der Praxis wäre hierfür ein Corpus notwendig der sämtliche Themengebiete und Sprachniveaus umschließt. Was allerdings wieder zur Folge hat, dass die Wahrscheinlichkeiten stark abweichen. Alternativ könnte eine Language Base für einen jeweiligen Bereich (Themengebiet und Sprachstil) geschaffen werden, dabei ist der Aufwand aber nur verlagert und ich müsste den Inhalt des Textes kennen, wofür ich vorher zumeist die Sprache kennen muss. Auch hier zeigen sich dieselben Ergebnisse, wie bei Tokenbigrammen.

Der Einsatz von Characterbigrammen ist aus der theoretischen Sicht sehr sinnvoll. Die Besonderheiten der Sprachen finden sich vor allem in zumindest Bigrammen oder besser in n-Grammen. Für das Englische wären z.B. *ly*, *ing*, *ed*, *ough*, etc. relevant, wohingegen im Deutschen Bigramme/n-Gramme, wie *ss*, *tz*, *sch* eine größere Rolle spielen. Auch [1] zeigte, dass die Verwendung von größeren Bigrammen eine höhere Genauigkeit bei der Vorhersage der Sprache bietet. Dies setzt allerdings voraus, dass die Trainingsdaten/der Corpus groß genug ist, ansonsten besteht das gleich Problem, wie schon bei den Token(-n-Grammen). Die Anzahl der Matches ist zu klein und die Vorhersage wird ungenau oder gar falsch. Tokenbigramme zeigten auch im praktischen Test gute Ergebnisse sowohl mit der Minkowski L1 Distanz als auch mit der Euklidischen Distanz. Die Kosinus Distanz zeigte keine guten Resultate und ist meiner Meinung nach für diesen Anwendungsfall eher ungeeignet.

4 Quellen

- Thomas Gottron und Nedim Lipka, A Comparison of Language Identification Approaches on Short Query-Style Texts, 2010. URL: `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.155.5553&rep=rep1&type=pdf`, aufgerufen am 19.11.2017