**Web-Mining Übung 1**

1. Verfahren um herauszufinden ob eine Hunderasse beliebt oder unbeliebt ist. Hierzu wird ein Crawler benötigt, der nach Seiten sucht auf denen die angegebene Hunderasse erwähnt wird. Dann müssen die gefundenen Erwähnungen in positiv und negativ klassifiziert werden. Hierzu wird dem Klassifizierer eine Datei mit positiven und negativen Beispielen bzw. Schlagworten die eine Klassifizierung zulassen (Bsp. Für negative unfreundlich, beißen, gefährlich…), zur Verfügung gestellt mit Hilfe dieser er die Klassifikation vornimmt. Wenn nun alle Daten klassifiziert wurden wird die Anzahl der positiven Einträge den negativen gegenübergestellt. Wenn es mehr positive Einträge gibt ist eine Rasse beliebt und wenn es mehr negative gibt unbeliebt. Es werden zusätzlich noch die Beliebtheit in % ausgeben also (positive Erwähnungen / Gesamtzahl der Erwähnungen).

Einsatz in der Praxis Eingabe einer Hunderasse => Ausgabe beliebt oder unbeliebt und Beliebtheit in %

Die Trainingsdaten werden müssen dem Klassifikator vorgegeben werden dies kann durch eintragen von Schlagworten in eine Textdatei erfolgen oder durch eine Abfrage von Beispielen der aktuellen Suche. Die Abfrage von weiteren Beispielen wird vor der Klassifikation durchgeführt.

**Positive Worte**

- Umgänglich

- Verträglich

- Lieb

- Kinderlieb…

**Negative Worte**

- Aggressiv

- Unfreundlich

- Beißen

- …

1. Die 30 am häufigsten Vorkommenden Worte sind bei den Texten ähnlich, somit würden die Texte bei einer Klassifizierung über die 30 häufigsten Worte wahrscheinlich fast alle Texte einer Klasse zuordnen. In der Tabelle ist dies auch zu sehen. In dieser sind die 30 Häufigsten Worte von zwei Texten sortiert nach Häufigkeit aufgezeigt, wie man sieht gibt unterscheiden diese sich kaum. 21 (rot markiert) der 30 Worte kommen in beiden unter den Top 30. Somit sollten diese für eine Klassifizierung ignoriert werden. Dies geschieht mit Hilfe von Stopp Words.

|  |  |
| --- | --- |
| **Text Geschwister** | **Text Jederman** |
| **ich** | und |
| **und** | Ich |
| **nicht** | jedermann |
| **sie** | Ist |
| **Du** | Die |
| **marianne** | Nit |
| **Mir** | Der |
| **fabrice** | Ein |
| **wilhelm** | Das |
| **Ist** | Mir |
| **S** | In |
| **Er** | Du |
| **Es** | Zu |
| **So** | Auf |
| **der** | mich |
| **wenn** | Mit |
| **mich** | Dir |
| **Die** | So |
| **das** | mein |
| **Zu** | Von |
| **ein** | was |
| **was** | daÃŸ |
| **den** | dich |
| **daÃŸ** | An |
| **wie** | Den |
| **In** | Wie |
| **sich** | sein |
| **Ihn** | Er |
| **mit** | Ihr |
| **Dir** | Sie |

1. Nach dem die Stoppwörter aus dem Ranking entfernt wurden,

Gibt es nur noch 6 Übereinstimmungen in den 30 häufigsten

Wörtern. Dies lässt würde bei einer Klassifizierung der beiden

Texte dazu führen das sie Wahrscheinlich in zwei verschiedene

Klassen sortiert werden.

|  |  |
| --- | --- |
| **Text Geschwister** | **Text Jederman** |
| marianne | jedermann |
| fabrice | nit |
| wilhelm | vetter |
| s | gesell |
| bruder | werke |
| immer | recht |
| wohl | geht |
| liebe | ja |
| muÃŸ | tod |
| sagen | gott |
| manchmal | mutter |
| herz | glaube |
| ganz | gar |
| gut | o |
| ab | wohl |
| gar | geld |
| nein | muÃŸ |
| glÃ¼cklich | kommt |
| weiÃŸ | jedermanns |
| hast | mann |
| recht | weiÃŸ |
| geht | wÃ¤r |
| wÃ¤re | hast |
| nie | schon |
| gern | buhlschaft |
| lieb | dicker |
| leben | halt |
| schon | weh |
| mehr | alls |
| lieben | stund |

1. a) Die Betrachtung wird ohne Stoppwörter gemacht!! für die Graphiken mit Stoppwörter siehe Ordner (word\_with\_stopwords)

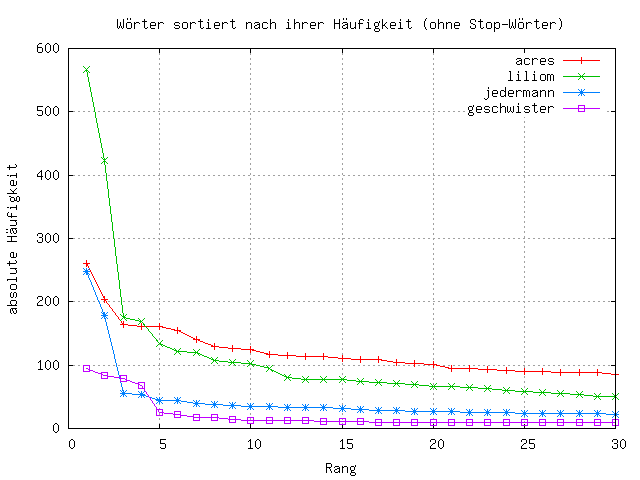


Abbildung 1 web-mining\ex1\results\word\plot\_abs.png

Bei der absoluten Verteilung kann man sehen, dass die ersten 5 Ränge die Skalierung so beeinflussen dass man in den darauf folgenden Bereich kaum Verlaufsunterschiede erkennen kann. Man kann aber hier ganz gut die Unterschiede im Bereich der ersten Ränge erkennen. Vor allem können aber verlaufsunabhängige Faktoren erkannt werden wie der Abstand wie die Start Werte und der Abstand zwischen den Kurven. Bei der Log Skalierung siehe Unten sind die unterschiedlichen Verläufe besser erkennbar vor allem auch im Bereich nach Rang 5.

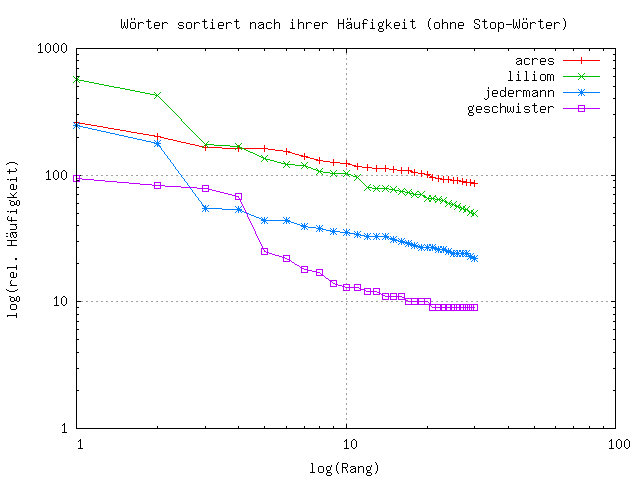


Abbildung 2 web-mining\ex1\results\word\plot\_abs\_log\_xy.png

b) Bei dieser Darstellung der Häufigkeiten der Wörter kann gut erkannt werden, dass die Verläufe zu Beginn sehr ähnlich sind und nur wenig schwanken dann nach ca. dem 7 Rang die Graphen eine hohe Varianz ausweisen und auch sehr unterschiedliche Ausschläge haben.

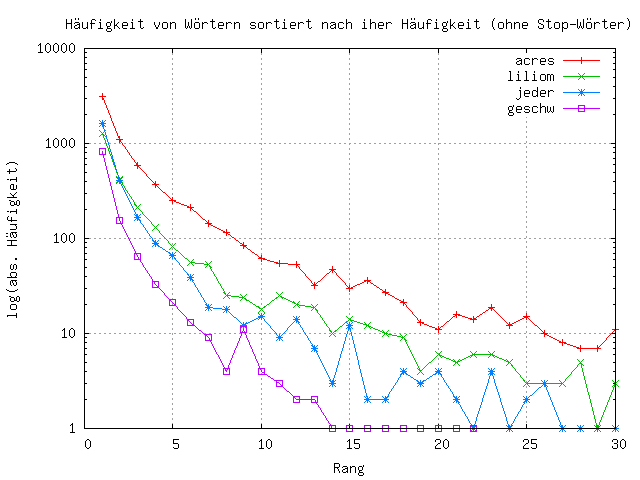


Abbildung web-mining\ex1\results\char\plot\_abs\_log.png

1. Bei den Texten in gleicher Sprache liegen deren Buchstabenhäufigkeiten dichter beisammen wie bei unterschiedlichen Sprachen was durch aus zu erwarten war. Es ist gut zu erkennen, dass es einige Buchstaben gibt die in den Sprachen sehr häufig vorkommen und andere die selten Verwendung finden. Aber auch dies Variiert von Sprache zu Sprache. Trotz gleicher Sprache unterschieden sich aber die Häufigkeiten in den Texten geringfügig.

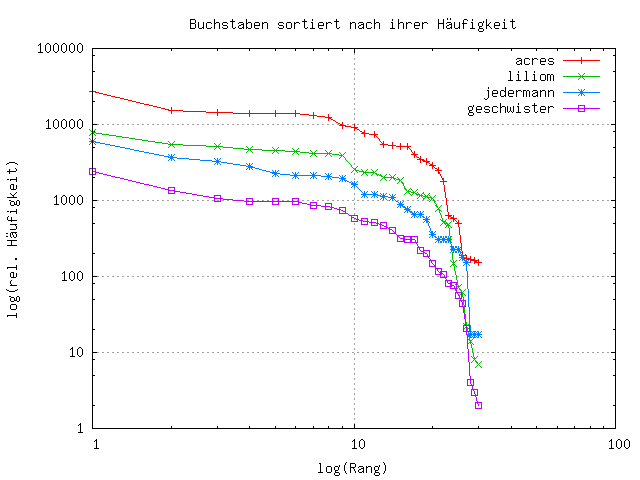


Abbildung web-mining\ex1\results\char\plot\_abs\_log.png

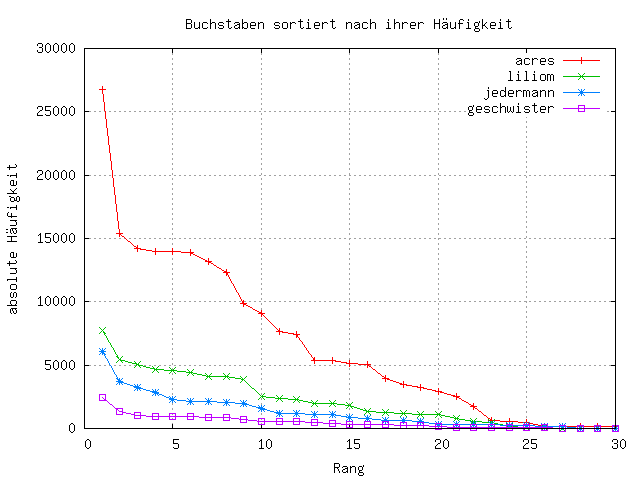


Abbildung web-mining\ex1\results\char\plot\_abs.png