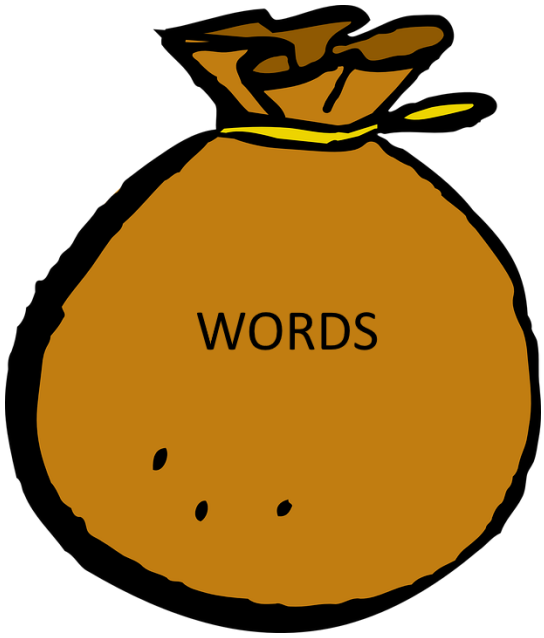
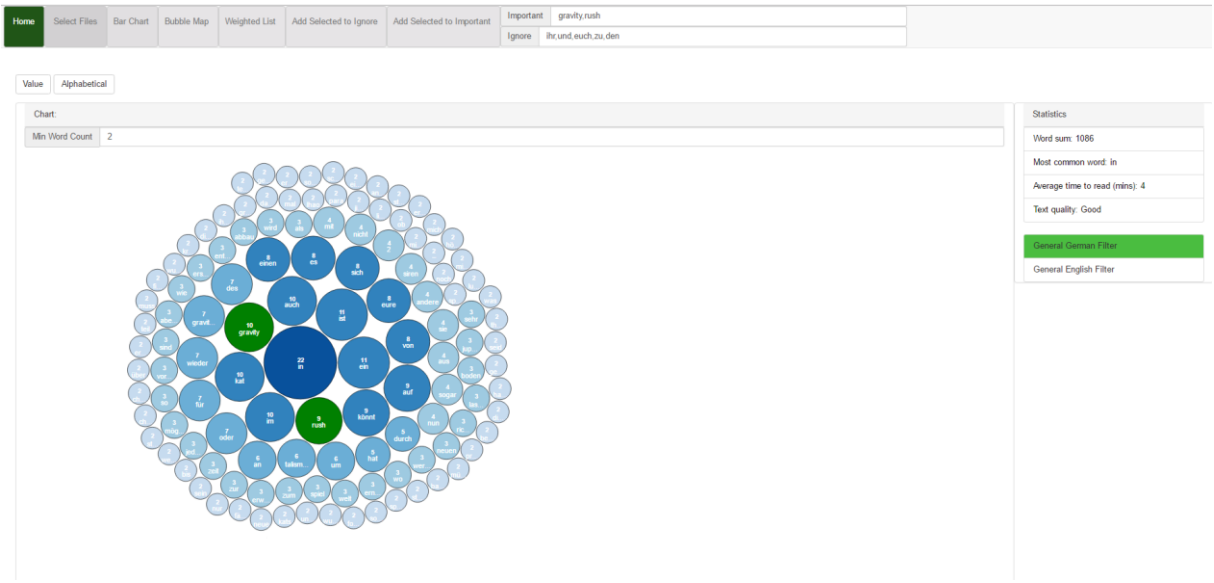
Bag of Words Visualization:



Team:
Benedikt Zöchling aka Bing (a1207430@unet.unvie.act.at)
Roman Karaba aka Bong (a1301624@unet.univie.ac.at)

Link zur Projektseite: homepage.univie.ac.at/a1301624/



Motivation und Problembeschreibung:

Motivation für unser kleines Visualisierungsprojekt war in erster Linie die Tatsache, dass Benedikt Zöchling ein Tool dieser Art gerne gehabt hätte, aber nirgendwo finden konnte. Eine weitere Motivation lag in dem zur Zeit der Themenwahl gerade auf Hochtouren laufenden US-Wahlkampf und weil bei beiden Teammitgliedern das Interesse bestand, im großen Stil überprüfen zu können, ob der mittlerweile offiziell gewählte 45. Präsident der vereinigten Staaten tatsächlich derartig oft das Wort „China“ verwendet wie ihm nachgesagt wird und wie groß der Anteil im Vergleich zu anderen auftretenden Wortwiederholungen ist. (Wie sich herausstellt sehr hoch).

Neben dieser eher humoristischen Anwendung, ist die Hauptanwendungsmöglichkeit unserer Tools aber natürlich die Textanalyse zur Verbesserung des eigenen Wortschatzes, bzw. des eigenen Schreibstils, sowie zur Analyse anderer Schreib- und Sprachstile. Der politische Ansatz wurde teils auch durch den Themen-Input von Herren Sedlmair inspiriert.

Das Tool ist dabei sprachunabhängig, dynamisch und vielseitig verwendbar und verlangt von seinen Nutzern keinerlei langwierige Eingewöhnung bzw. irgendeine Form von Fachkenntnis. Ein Tool für Jedermann sozusagen.

Daten:

Grundgedanken des Tools ist es, dass der User keinerlei spezifische formatierten .csv-Dateien oder Ähnliches benötigt um es zu verwenden, sondern einfache Text-Dateien. Es muss dabei auf nichts besonders geachtet werden und es können auch alle zu analysierende Texte in ein File geworfen werden. Sein volles Potential erreicht das Tool aber bei der Verwendung von mehreren Textfiles. Sollte man eine chronologische Analyse betreiben wollen, müssen die Files lediglich in der chronologisch korrekten Reihenfolge hochgeladen werden. Wir achten dabei betont nicht auf Änderungsdaten oder verlangen besondere

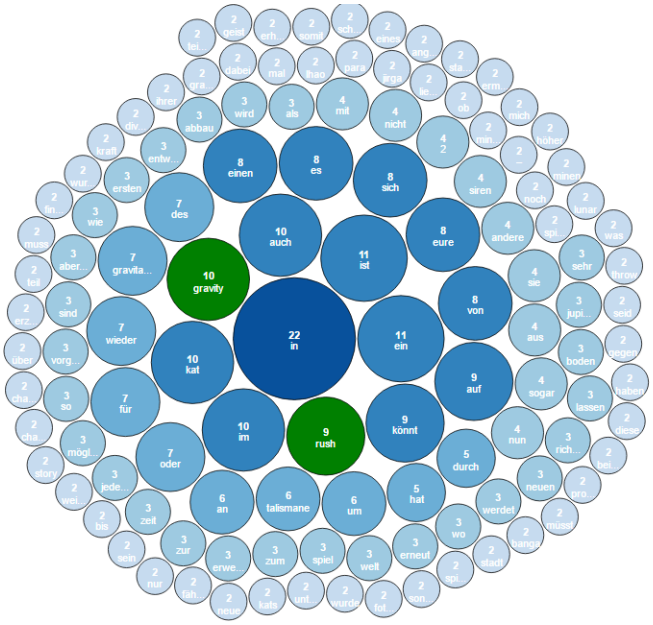
Namensformatierungen, da nichts simpler ist, als einfach die Files in der gewünschten Reihenfolge anzuklicken. Alles in diesem Tool ist

Dateinamen um dem User möglichst wenig in seinen Möglichkeiten einzuschränken. Komfort führt oft dazu, dass Tools nur für einen spezifischen Anwendungsfall geeignet sind oder kaum alternative Anwendungsmöglichkeiten geboten werden. Unser persönlicher Reiz liegt aber darin ein Tool zu erschaffen, dass möglicherweise in Anwendungsgebiete genutzt wird, an die wir selbst nicht einmal gedacht haben und deswegen verlangen wir von unseren Nutzern nichts weiter als Texte zur Eingabe in einer beliebigen Form.

Die Texte werden nach dem Hochladen in klassischer Bag of Words-Form durchgelaufen, es erhält also jedes vorkommende Wort einen Eintrag in unserem .json mit einem Wert der mit jedem Weiteren vorkommen desselben Wortes um eins erhöht wird. Dies wird für jedes File getrennt gemacht, sodass die Werte je nach verlangen zusammen addiert oder getrennt betrachtet werden können. Weiter werden die entstehenden Daten sofort analysiert. So wird der gesamte Wordcount, das meist verwendete Wort, eine durchschnittliche Lesezeit sowie ein Hinweis auf die Qualität des Textes gegeben (dazu später mehr).

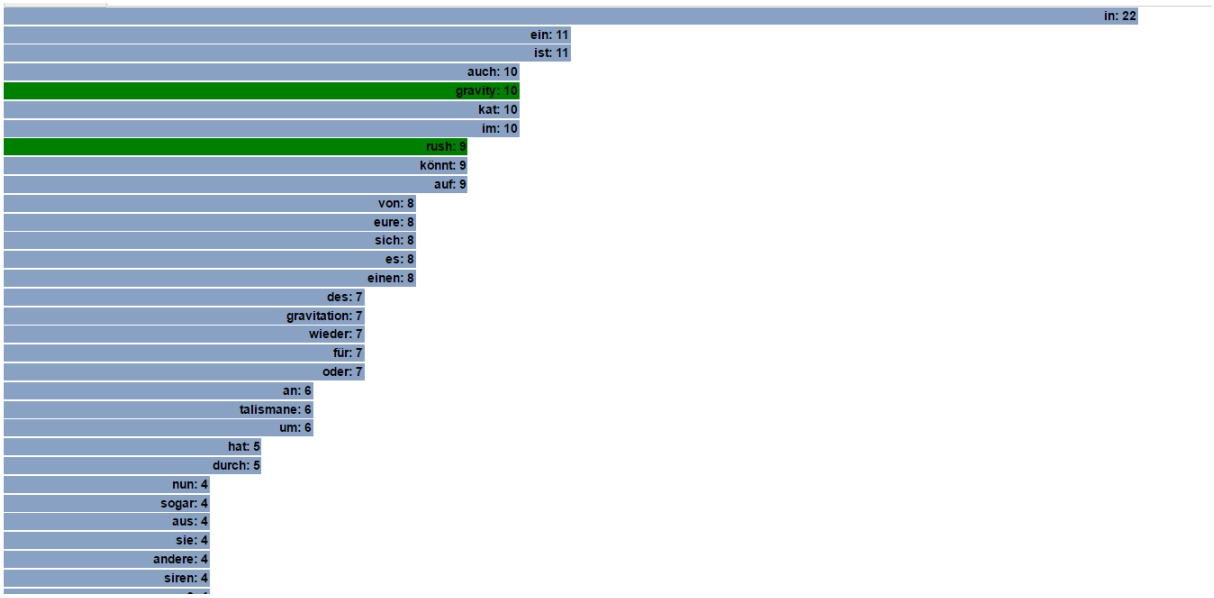
Visualisierung (Approach):

Für die Visualisierung der so entstehenden Daten haben wir uns für drei Diagramme entschieden, zwischen denen je nach Anwendungsfall dynamisch gewechselt werden kann. Das dynamische Wechseln hat dabei den Vorteil, dass jedes einzelne von ihnen in bildschirmfüllenden Ausmaßen vorhanden sein kann. Dashboards sind wundervolle Werkzeuge um in einem Blickfeld so viele Informationen wie nur irgendwie möglich unterzubringen. Sie haben aber oft eine leicht „erschlagende“ Wirkung auf den Nutzer, vor allem wenn es darum geht Daten mit tausenden Einträgen zu analysieren. Unser Tool soll aber jedem Menschen sein volles Potential zugänglich halten, ohne ihn bereits in der ersten Ansicht zu übermannen. Daher haben wir uns entschieden unsere drei Diagrammtypen in unterschiedlichen Ansichten zu positionieren. Dies ergibt vor allem deswegen Sinn, da zwei unserer Diagramme einen jeweils differenten Ansatz haben eine „ein Blick“ Lösung zu bieten, währen das dritte Diagramm dafür geeignet ist, spezifische Vergleiche anzustellen bzw. in die Drill-Down Funktion zu bieten.



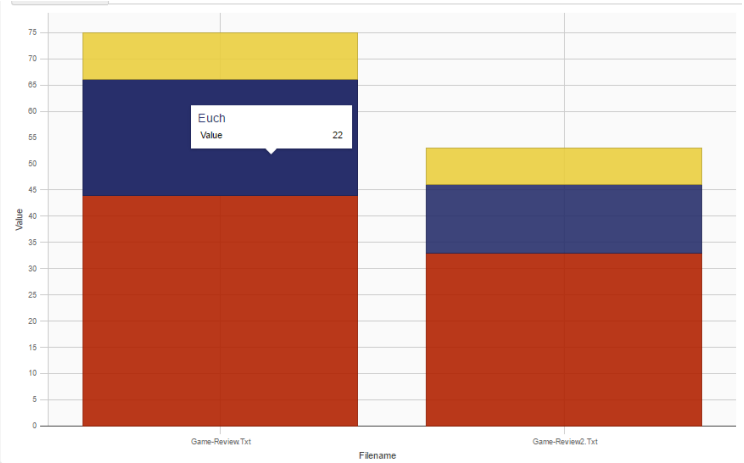
Die Bubble Map:

Der Name Bubble-Map mag möglicherweise nicht wirklich existent sein, dennoch fanden wir ihn passend wegen der Google Maps-artigen Funktionalität die in das klassische Bubble-Chart integriert wurde. Diese ist deswegen so wichtig da eine normale Bubble-Chart im Bereich der kleineren Blasen meist nur noch eine Wolke aus winzigen Bläschen darstellt, hier also für User die sich vielleicht auch dafür interessieren, welche Wörter nur zwei bis dreimal in einem Textvorkommen, wieder eine Einschränkung entstehen hätte können. Diese können dann in den Außenbereich des Charts zoomen und sich in diesem entlang ziehen um so auch die Relationen der kaum verwendeten Worte miteinander vergleichen zu können. Der klassische Anwender kann aber auf einen Blick erkennen, bei welchen Worten möglicherweise eine Problemzone entstanden sein könnte, welche Worte besonders oft im Text vorgekommen sind wie die allgemeine Verteilung der Worte sich präsentiert. Prinzipiell ist es erstrebenswert eine Bubble-Map zu erzeugen, die eine möglichst große Anzahl von verschiedenen Blasen mit möglichst gleicher Größe aufweist. Sortiert man mit der Ignorier-Funktion oder mit einem der vorgefertigten Filter, welche ich beide später noch genauer erklären werde, unwichtige Wörter wie Artikel aus der Bubble-Map, sollte diese möglichst wenige Differenzen zwischen den Blasen anzeigen. Dominieren einige wenige Blasen klar das Bild, könnte es sich hier um eine Problemzone oder zumindest eine interessante Auffälligkeit handeln.



Die weighted List:

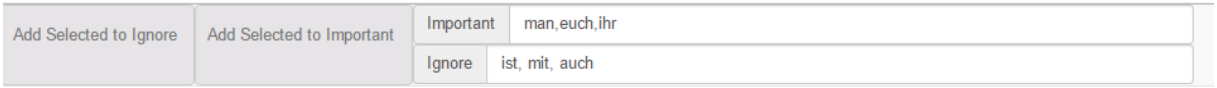
Im Grunde handelt es sich hier um ein horizontales Bar-Chart dessen Vorteil jedoch sein Listen-Charakter ist. Diese Funktion wurde als zusätzliche allgemeine Ansicht auf Raten von Herrn Sedlmair hinzugefügt und bietet einige klare Vorteile gegenüber der Bubble-Map. So lassen sich Größen-Relationen zwischen den Wörtern viel klarer erkennen und vor allem spezifische Problemwörter schneller identifizieren. In den Beispiel oben sieht man als Veranschaulichung wie das Wort „in“ tatsächlich klar die Liste dominiert und die im Ranking nachfolgenden Worte „ein“ und „ist“ nur halb so oft vertreten sind. Etwas das bei der Bubble-Map von vorhin nicht so deutlich zu erkennen war. Weiter ließ sich die Liste mit der Zusatz Funktionalität ausstatten, sie außer nach dem Vorkommen der Wörter, auch nach dem Alphabet zu sortieren. Bisher ist uns leider keine Anwendung eingefallen bei der dieses Sortierverfahren tatsächlich nützlich wäre, aber möglicherweise möchte jemand einen Text verfassen indem möglichst wenige Wörter die mit „a“ beginnen vorkommen und hat dann mit unserem Tool genau das richtige zur Hand um die „Qualität“ seines Textes zu überprüfen. Da dies unserem Ansatz ein Tool für „alle“ zu erschaffen entsprach, haben wir diese Funktion also eingebaut. Vorteil der Bubble-Map gegenüber der Liste ist übrigens deren Kompaktheit. Würde jemand beispielsweise Texte mit insgesamt 1000 verschiedenen Wörtern hochladen, könnte er bei der Bubble-Map dennoch in Sekunden schnelle die als Important markierten Wörter (dazu später mehr) finden, während die weighted List keine Möglichkeit bietet, diese weiter zu komprimieren und somit Minuten langes scrollen verlangen könnte. Da wir beide Optionen anbieten, kann der User somit einfach seine präferierte Ansicht wählen.



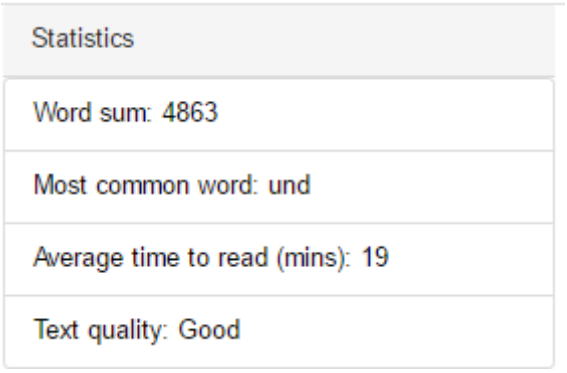
Das Stacked Bar-Chart:

Sozusagen die Drill-Down-Funtion unseres Tools. Sowohl in der Bubble-Chart, als auch in der Weighted-List ließen sich per Klick mehrere Wörter auswählen, die nun weiterführend überprüft werden. Diese können anschließend in einer Stacked Bar-Chart angesehen werden. Das Besondere daran ist, dass ihr vorkommen nach allen ausgewählten Files sortiert wird. Eben das ermöglicht dann eine chronologische Analyse oder eben einen direkter Vergleich zweier Autoren anstellen. Es wäre aber auch möglich Arten von Schreibstilen, Aggressivität von politischen Gegnern oder die Häufigkeit von verwendeten Fremdwörtern bei zwei Sprachen zu vergleichen.

Andere Funktionen:

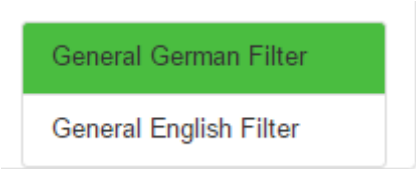


Ignoreliste: Simpel und zugänglich, kann sich jeder User in jeder Sprache die er möchte eine Liste von Wörtern zusammenstellen die ihn einfach nicht interessieren, woraufhin diese in keinem der Diagramme mehr aufscheinen.
Important: Zur erleichterten Suche von spezifischen Worten auf die man achten möchte, gibt es eine ebenso zugängliche Importantliste. Alle hier aufgelisteten Worte werden egal wie oft sie vorkommen in einem knalligen klar ersichtlichen Grün hervorgehoben. Dies ist vor allem in der Bubbel Map immens hilfreich, da in den kleineren immer blasser werdenden Bereichen, die natürlich immer dichter an Blasen werden, selbst ein kleiner in sattem Grellgrün leuchtender Punkt heraussticht.
Add-Buttons: Als Alternative zu dem manuellen Schreiben können in der Bubble-Map als auch der Weighted-List einfach mehrere Wörter angeklickt und anschließend zu einer der Listen hinzugefügt werden.



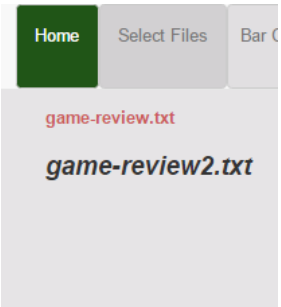
Statistiks:

Ein simpler Komfort für einige mögliche Anwendungen. Möchte man z.B. einfach wissen wie viele verschiedene Worte in dem Text vorkommen, oder schnell das am häufigsten verwendete Wort erfahren, hat man diese Informationen hier auf einen Blick. Außerdem wird errechnet wie lange das Lesen des Textes in etwas dauern würde, gemessen an durchschnittlich 250 Wörtern pro Minute, was beispielsweise für Präsentationen relevant sein kann. Zu guter Letzt wird noch Aufschluss über die Qualität des Textes gegeben. Diese kann „bad“, „ok“ oder „good“ sein und errechnet sich aus zwei Faktoren: Erster Faktor: Wie viele Worte mit dem seltensten Wortvorkommen gibt es. Wenn die Summe dieser über 50% des gesamten Textes ausmacht, wird dies als gut erachtet. Zweiter Faktor: Wenn das am häufigsten vorkommende Wort weniger als 5% des gesamten Textes ausmacht, wird dies als gut geachtet (Wert sinkt bei besonders kurzen Texten ein wenig).



Vordefinierte Filter:

Sind prinzipiell zwar nicht nötig, da sich jeder seine Filter selber generieren kann, dennoch haben sie natürlich einen gewissen Komfort-Faktor. Drückt man auf die Taste wird der Ignore-Liste automatisch ein Set von in den meisten Fällen nicht relevanten Wörtern wie Artikeln hinzugefügt. Diese Option ist jedoch mit Vorsicht zu genießen und wird derzeit auch nur in English und in Deutsch angeboten.



Selected Files:

Eine Liste aller hochgeladener Files, die es ermöglicht einzelne Dateien zu ignorieren um beispielsweise einen kurzen Blick auf eine einzelne von ihnen zu werfen oder jeweils zwei von vielen im Detail zu vergleichen.

Anwendungsbeispiele:
User:

Name: Benedikt Zöchling
Arbeit: Freier Redakteur/ Autor
Bei: Shock 2 und Anderen
Alter: 24
Typ: Gewinner

Beschreibung:
Das ist Benedikt, außer dass er gerne Steaks isst, ist er außerdem freier Redakteur und Autor von dämlichen Kurzgeschichten. Benedikt schreibt sehr gerne, war aber leider nie so richtig gut darin, weswegen Benedikt diesen Umstand gerne verbessern möchte. Aber wie nur ?

Anwendung:
Natürlich mit dem Bag of Words-Visualization Tool von Team Bing and Bong. Indem Benedikt einige seiner bisherigen Spiele-Reviews in einer Textdatei zusammenfasste, sich eine Liste mit für ihn irrelevanten Wörtern erstellte(die er gleich in derselben Datei speicherte) und alles anschließend hochlud, fiel Benedikt durch einen Blick auf die Bubble-Map auf, dass er die Worte „man“, „euch“ und „ihr“ ziemlich oft verwendet. Nun hätte es aber sein können, dass diese Worte in Spiele-Reviews einfach nicht zu vermeiden sind. Um dies zu überprüfen, erstellte Benedikt eine weitere Text-Datei, in die er Reviews von einem Autoren steckte, den Benedikt sehr bewundert. Er schaute, dass die beiden Dateien möglichst gleichstark befüllt sind und lud sie hoch. Anschließend, trug er in die Important-Liste die gesuchten Worte ein und konnte sie so im Handumdrehen markieren. Als er sich dann in der Stacked-Barchart den Vergleich ansah, fiel Benedikt auf, dass sein Vorbild besagte Wörter in seinen Reviews offensichtlich nur halb so oft verwendet wie er selbst. Daher fasste Benedikt den Entschluss, zukünftig mehr darauf zu achten diese Wörter eher zu vermeiden.

Einige Zeit später, stellte sich Benedikt die Frage, ob er seine Verbesserungspläne denn nun wirklich umsetzen konnte. Also fasst er wieder eine Datei mit aktuelleren Texten von sich zusammen, lud sie zusammen mit den beiden vorherigen Dateien hoch und verglich die relevanten Wörter wieder in der Stacked-Barchart. Dabei erkannte er, dass er diese problematischen Wörter nicht nur deutlich weniger oft als früher verwendete, sondern sogar ziemlich ähnliche Wortanzahlen wie sein Vorbild erreichte. Ein Erfolg auf ganzer Linie. Dank des Bag of Words Visualization-Tools von Team Bing and Bong fiel es Benedikt also leicht, sein bisheriges Schreibverhalten zu analysieren und weiterführend zu verbessern.

Potentieller User:

Name: Mrs. Hillary
Arbeit: wäre gerne Präsidentin
Bei: Amerika
Alter: alt
Typ: Nicht-Gewinner

Beschreibung:
Hillary wäre sehr gerne Präsidentin der vereinigten Staaten von Amerika. Leider wird sie immer wenn sie es versucht, von ihrer Konkurrenz ausgestochen. Hillary möchte nun herausfinden, wieso dies nun schon wieder passiert ist. Aber wie nur ?

Anwendung:

Natürlich mit dem Bag of Words-Visualization Tool von Team Bing and Bong. Hillary fasste einige Mitschriften ihrer eigenen politischen Reden und der Reden ihres Konkurrenten in jeweils eine Textdatei zusammen und lud sie in das Tool. Als sie sich die beiden Dateien dann jeweils separat ansah, fiel ihr gleich als erstes auf, dass ihr Konkurrent nicht einmal halb so viele verschiedene Worte verwendet hatte wie sie. Auch viel ihr auf, dass einige ganz spezifische Worte von ihrem Konkurrenten unfassbar oft verwendet wurden.

Also markierte sie besagte Worte und wechselte in die Stacked-Barchart um zu sehen, wie oft sie selbst diese Worte im Vergleich verwendet hatte. Dabei fiel ihr auf, dass sie besagte Worte nur einen Bruchteil so oft verwendete wie ihr Konkurrent und einige von ihnen bei ihr überhaupt nur ein bis zwei Mal erwähnt wurden. Also scheint es so, als ob die Wähler eher einen Kandidaten wählen, der wenige Worte möglichst häufig benutzt. Dank des Bag of Words Visualization-Tools von Team Bing and Bong, weiß Hilary dies nun und kann es beim nächsten Mal besser machen.

Potentieller User:

Name: Mr. Serious Buisness
Arbeit: macht Serious Buisness
Bei: einem seriösen Unternehmen
Alter: 40
Typ: Serious

Beschreibung:
Mr. Serious Buisness möchte nichts an sich selbst verbessern, denn er macht Serious Buisness und er macht es gut. Leider hat sein seriöses Unternehmen öffentlich einen eher negativen Ruf und viele meinen es wäre Skruppel- und Seelenlos. Um sich die ständigen Obstanschläge auf sein Auto und damit zusammenhängende Waschstraßen-Besuche zu ersparen, möchte Mr. Serious Buisness einen unbezahlten Praktikanten anstellen, der auf den Social-Media Plattformen die Wogen glättet. Er hat zwei Bewerber und möchte herausfinden wen er anstellen soll. Aber wie nur ?

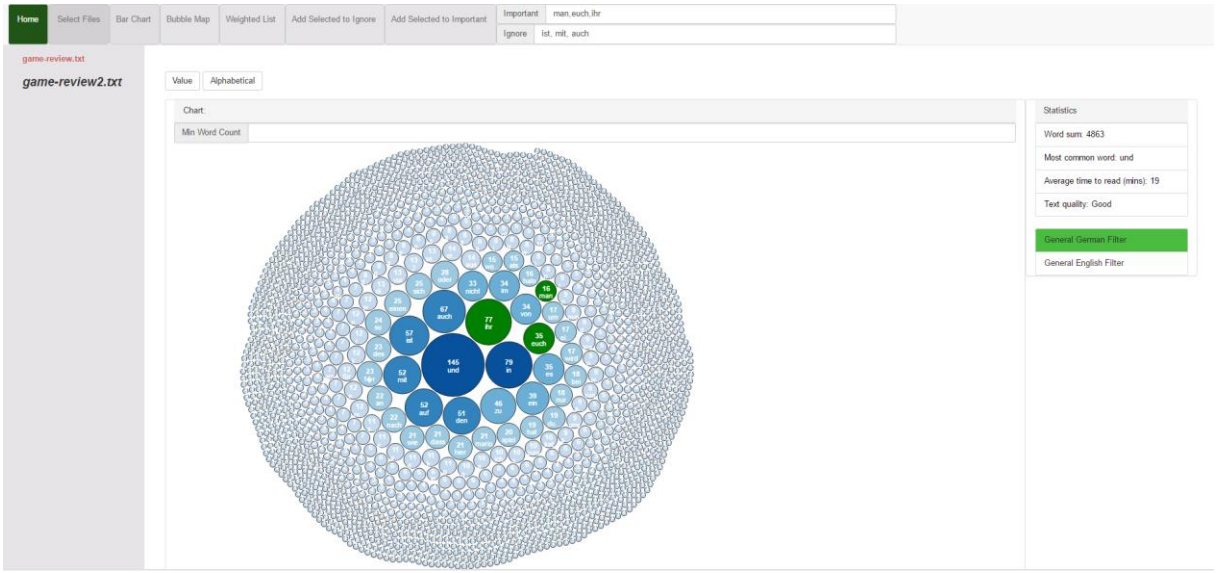
Anwendung:

Natürlich mit dem Bag of Words-Visualization Tool von Team Bing and Bong. Mr. Serious Buisness kopiert einfach die zu den Bewerbungen hinzugefügten Beispieltexte in jeweils eine Datei. Anschließend lässt sich Mr. Serious Buisness eine große Textdatei mit allen Whatsapp Unterhaltungen der größten Anti-Serious-Buisness Redelsführer anfertigen. Alle drei werden dann in das Tool hochgeladen.

Zuerst werden in der Weighted Liste einfach die ersten 20 am häufigsten verwendeten Worte markiert, dann wird in die stacked Barchart gewechselt wo Mr. Serious Buisness ganz leicht vergleichen kann, welcher der Bewerber den Tonfall der Redelsführer besser erwischt. Dank des Bag of Words Visualization-Tools von Team Bing and Bong, weiß Mr. Serious Buisness nun ganz genau, welchen Bewerber er anstellen soll und kann wieder seinem Serious Buisness nachgehen.

Implementierung:

Umsetzung erfolgte mittels Javascript für die Erstellung des Bag of Words, sowie einiger Funktionen der Homepage sowie d3 bzw. d3plus für die Diagramme. Bei der Weighted-List haben wir d3 ohne ein dazugehöriges SVG verwendet um das Verwenden von SVG internen Scrollbalken zu verhindern und weil wir es außerdem interessant fanden, wie weit wir mit diesem Ansatz kommen können. Für die Bubble-Map haben wir das klassische SVG genommen, da das Map-Verhalten sich natürlich ohne SVG nur schwerlich umsetzen lässt.



Probleme:

Da wir beide nebenbei Berufstätig sind, war Zeit ohne Frage das größte Problem und hat uns bis zu dem jetzigen Zeitpunkt davon abgehalten, wirklich alle Funktionalitäten umzusetzen die wir uns so vorgestellt hatten. Wir haben aber vor, dieses Tool auch über dieses Semester hinaus noch weiter zu verfeinern und mit neuen Funktionen zu erweitern, da wir es wie bereits erwähnt zu einem nicht unwesentlichen Teil für uns selbst gemacht haben. Andere Probleme gab es unter anderem dabei, dass wir uns lange Zeit nicht sicher waren ob wir d3.v3 oder d3.v4 verwenden wollen.

Feedback:

Aus dem Feedback das wir bisher sammeln konnten kristallisierten sich für uns folgende Stärken und Schwächen unseres Tools heraus:

Stärken:

- Dynamisch
- Sprachunabhängig (beispielsweise für Ungarisch)
- Vielfältige Anwendungsmöglichkeiten
- Verlangt keine Vorkenntnisse
- Verlangt keine besondere Formatierung

Schwächen:

- Kein direkter Textinput statt Dateien
- Keine Möglichkeit zu Speichern
- Keine Möglichkeit zusätzliche Dateien hochzuladen, wenn bereits welche geladen sind
- Bubble-Map teilweise auch mit Zoom nicht gut genug lesbar.

Geplante Verbesserungen:

Wie bereits erwähnt, haben wir uns an sich vorgenommen, auch weiterhin an unserem Tool zu arbeiten, da wir einiges aus Zeitmangel nicht umsetzen konnten. Pläne dafür sind:

- Weighted List mit einer Dateien-Liste daneben, welche automatisch alle Worte markiert, die in der Datei vorkommen wenn man sie anklickt.
- Bessere Filter für weitere Sprachen
- Optisches Makeover der gesamten Seite
- Text-Qualität-Tool präzisieren
- Verbesserung aller oben genannten Schwächen.

Lessions Learned:

- D3 ist ein überraschend vielfältiges Tool auf das wir möglicherweise auch bei zukünftigen Projekten noch zurückgreifen werden.
- Javascript kann teilweise recht praktisch sein.
- .jsons sind toll
- Visualisierung bietet tolle Möglichkeiten sich den eigenen Alltag zu erleichtern.

Performance:

Läuft überraschend flott auch bei großen Textdatein. Der Sprung zwischen den Charts ist ebenfalls mit kaum merkbarer Verzögerung möglich. Kleinere Bugs beim Zoomen ließen sich bisher nicht ausmerzen.

Related Work:

Wie Anfangs erwähnt ist dieses Tool teils aus dem Problem entstanden, dass so ein Tool für uns nicht auffindbar war. Einige Inspiration ist allerdings von der Seite https://wordcounter.net/ gekommen, die zwar keine direkte Visualisierung anbietet, aber relativ sinnvolle und interessante Statistiken anführt.

A clear separation of tasks between the group members:

Roman Karaba : Erstellung der Tabellen aus den Textdokumenten, Upload der Textdokumente. Seperierung nach Ignore, Important, und Filename. Statistiken.
Benedikt Zöchling: Bubble-Chart, Bar-Chart, Weighted List, Homepage Design und dieses File hier.
Gemeinsam: Planung und Unterstützung durch Extreme Programming.

References:

Die Design Entscheidungen basieren auf Ideen von:
T. Munzner: Visualization Analysis & Design: Abstractions, Principles, and Methods,
CRC Press, 2014

Ideen teilweise von der d3 Example Seite:
http://techslides.com/over-1000-d3-js-examples-and-demos