OneTIPP Algorithmus

Version: 23.11.2015 v3.0
Bearbeiter: Sebastian Enger
Status: Geheim

Die Vision von OneTIPP für das Jahr 2016

1. Ziel: Produktion von neuen Textinhalten softwaregestützt in Echtzeit mit nur einem Klick
2. Eingabetexte in Echtzeit zu verändern (Q1-2016):
    - Leistungsschutzkonform
    - Urheberrechtlich sicher
        - =>Ziel: Massenmarkt
    - Google „Duplicate Content“ konform (Q3-2016)
        - =>Ziel: Alle Menschen mit E-Commerce Ambitionen
3. Ständige Verbesserung des Algorithmus bis hin zum „Automatischen Maschinen Lernen“ und „Künstlicher Intelligenz“ (=Forschung & Entwicklungsauftrag) (Q4-2016):
    - K.I. soll aus den Texten lernen und sich langfristig selbst verbessern
    - Die Qualität der Ausgabe soll ständig weiter verbessert werden (=Ziel: Qualitätsführer)

Das Anforderungsprofil an die erste OneTIPP Version

1. Echtzeitverfügbarkeit (Q1-2016)
    - Eingabe, Verarbeitung und Ausgabe im Echtzeitbereich
2. Ausgabe von strukturierten, veränderten Texten
    - für den Massenmarkt (=Wachstum, B2C) (Q1-2016)
    - von professionellen Texte für Journalisten, Techniker und technische Dokumentationen (=B2B) (Q2-2016)
3. Unterstützung von Autorenprofilen (Q1-2016)
    - Die Engine soll anhand von definierten und selbstlernenden Logiken, Eingabe- und Clientanfrageparametern erkennen, wer welchen Schreibstyl hat
    - die Ausgabe des OneTIPP Textes an diesen Schreibstyl des Originalautors anpassen
    - Vorteil:
        - Der Schreibstyl des Autors bleibt erhalten
        - Er bekommt verschiedene Varianten seines Urheberwerkes
    - Markieren der Stimmung:
        - Aus einem traurigen Text soll kein Lustiger gemacht werden
    - Sprachlevel der Ein- und Ausgabetexte soll gleich sein
4. Stil und Grammatikveränderungen (Q2-2016)
5. Verbesserung von Schreibfehlern (Q2-2016) und Grammatik (Q4-2016)
6. Unterstützung von Transliteration (Q1-2016)
7. Automatische Übersetzungen (Q1-2016 – Deutsch > Englisch)
8. Multilingual und Global
    - Deutsch (Q2-2016)
    - Englisch (Q3-2016)
    - (später: Spanisch, Portugiesisch, Französisch, Italienisch, Russisch, Chinesisch)
9. Grundlegendes „Maschine Learning“ (Q1-2016)
    - zur Nachträglichen Analyse und ständigen Verbesserung des OneTIPP Algorithmus
    - Eventuell Tensorflow benutzen?
        - http://tensorflow.org/
        - https://github.com/tensorflow/tensorflow

10. Qualitätsprüfung nach dem Verändern (Q1-2016)

Ablauf des Veränderungsprozesses bei OneTIPP

1. Eingabetext in Webseite einfügen oder Public Web Service API ansprechen
2. Intern: speichern der Eingabetexte für spätere Bearbeitung und zur Nutzung zur Verbesserung der Algorithmen von OneTIPP
3. Verarbeitung durch den OneTIPP Algorithmus
4. Ausgabe des Textes auf der OneTIPP Webseite oder mittels Web Service
    a. Die Ausgabe auf der Webseite ist editier und exportierbar
    b. Die exportierte Version wird internen Datenbank gespeichert
5. Nachanalyse durch das OneTIPP Team zur Verbesserung des Algorithmus

Anmerkungen zur Entwicklung von OneTIPP

1. Professor Heyer hat Bedenken, dass wir keine Open Source Software nutzen dürfen, wenn wir später verkaufen wollen („Software frei von Rechten Dritter“)
    a. Gegenargumente:
        i. Zeitkritische Entwicklung („First come, First serve“)
        ii. Funktionalität erstellen und Prototyp entwickeln
        iii. Später: Wandel von Fremdbibliotheken hin zu Eigenentwicklungen, wenn das Team groß genug ist und entsprechende Ressourcen vorhanden sind
2. Konzentration auf Kernkompetenzen:
    a. Priorität:
        i. Einfache Texte für den Massenmarkt für B2C und Massenmarkt (=damit als Beta online gehen und Feedback sammeln)
        ii. Parallel dazu erste Studien und Forschungsarbeit starten, wie man professionelle Texte ausgeben kann (=B2B)
    b. „Kurz, knapp, funktional“:
        i. Die Vision soll umgesetzt werden
        ii. Kein Totprogrammieren bzw. kein massenhaftes Zufügen von unnötigen, weiteren Features
        iii. Es wird eingebaut und erweitert was die Qualität verbessert bzw. von Kunden gefordert wird (=was den Kundennutzen erhöht)

Mögliche Forschungsobjekte zur OneTIPP

1. Wie kann man, statt eines Einzelnen, gleich ganze Satz- und Textblöcke, verändern?
    a. Reduktion von komplexen Sätzen, Satzkombinationen oder Textblöcken in einfach strukturierte, vereinfachte Satzmuster
2. „Maschinen Learning“ bzw. „Künstliche Intelligenz“ zum Erkennen von Schreibstylen, Grammatik und Rechtschreibregeln aus bestehenden Texten
3. „Maschinen Learning“ bzw. K.I zur automatischen Erlernen von neuen Sprachmustern/Spracheigenheiten und neuen Sprachen (Russisch, Italienisch, Französisch, Chinesisch, Japanisch) zur Implementation in OneTIPP

Der OneTIPP Algorithmus im Detail

1. Lege ein Autorenprofil an
    a. Benutze dazu Hidden Markov Modelle

        i. „Sprachstrom gewisse Eigenschaften zuzuweisen“ (Heyer, Quasthoff, & Wittig, 2012), S. 115
        ii. „Vorhersagen über die Fortsetzung des Sprachstroms“ (Heyer, Quasthoff, & Wittig, 2012), S. 115
        iii. Speicherung der Wahrscheinlichkeiten der auftretenden Wortketten als Autorenprofilfingerabdruck
    b. Führe eine Stimmungsanalyse durch (Heyer G. , 2012), S. 7ff
        i. Was sind bekannte und beliebte positive oder negative Schlagwörter des Autors?
    c. (Speichere den POS-TAG Fingerabdruck des Autors – Welche Satzbaustrukturen werden gern verwendet?)
    d. Berechnung der Wahrscheinlichkeiten eines Eingabesatzes – n-Gramme und Bestimmung der „Maximum Likehood Estimation“ (MLE) (Heyer, Quasthoff, & Wittig, 2012), S. 102
    e. Berechne die Lesbarkeitsstatistiken des Eingabetextes

2. Führe ein POS-Tagging des Eingabetextes durch
    a. Aktuell verwende ich den TreeTagger der UNI Stuttgart als STTS Modell (http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/)
    b. Frage an Prof. Heyer: Welchen POS-Tagger für DACH empfehlen Sie?
    c. Speichere den POS-TAG Fingerabdruck des Autors

3. Definiere Logiken von STTS Tags, die nicht ausgetauscht werden dürfen
    a. z. B. „ART“, „ADV“ an Hand der POS-Tags des Eingabetextes
    b. ….

4. Definiere Logiken zum Zusammenfassen und für den Austausch
    a. Wandel „CARD“ (=Kardinalzahl) als Nummer bzw. Zahl im Orginaltext in ein Wort um, und umgekehrt (=https://pypi.python.org/pypi/num2words)
    b. Fremdwörter (POS-TAG=FM) ins Deutsche übersetzen und austauschen
    c. Zum Beispiel: „Ich [aß Eis]“ (PP-VV-NN)
        i. Nomen nur mit Nomen tauschen
        ii. Verb nur mit Verb tauschen

5. Sinnhafte Kernelemente jedes Satzes markieren
    a. Frage: Wie kann man „bedeutungstragende Wortformen bzw. Wortgruppen“ automatisch finden? -> Vgl. „normalisierten Termfrequenz“ (Heyer, Quasthoff, & Wittig, 2012), S. 204ff
    b. Der Sinn der Kernelemente eines Satzes muss erhalten bleiben (=Version 1), die Kernaussage des Textes muss erhalten bleiben (=Version 1)
    c. Die sinnhaften Kernelemente eines Textblocks soll sinnbehaltend verändert werden, wobei die Kernaussage des Textblocks gleich bleibt (=Version 2)
    d. Kernelemente sind Sinn-tragende Hauptelemente
        i. Zum Beispiel: „ich [aß Eis]“
        ii. Zum Beispiel: „ich [ging spazieren] und [traf Harry Potter]“
        iii. Logiken für Kernelemente bestimmen:
            1. Zum Beispiel: „Verb Nomen“, „Verb Adjektiv“, „Verb Adjektiv Nomen“
    e. Sinnhafte Kernelemente werden passend Sinnverwand getauscht (Synonym, Antonym):
        i. Zum Beispiel: „ich [aß ein Eis]“
        ii. Kernelement: VERB-POSSESIVPRONOMEN-NOMEN
        iii. Bestimme von „ich [aß ein Eis]“:
            1. Zeitform („Präteritum“)
            2. Person („1. Person Singular“)

3. Geschlechterbestimmung eines/des folgenden NOMEN
    a. [EIS] → männlich

iv. Austauschverhalten:
1. Nimm ein Verb aus der Synonym Datenbank und bringe es in die Zeitform, Person aus (Punkt 5. d. iii.)
2. Nimm ein Nomen aus Synonym DB und passe bzw. bringe es zu dem Geschlecht passend aus (Punkt 5. d. iii. 2.) dar

f. Bestimme die semantische Relation zwischen signifikanten Kookurenzen beim Austausch von Synonymen, Akronymen, Füllwörtern
    i. Tausch möglichst nur bei Kookurrenz mit möglichst hoher Signifikanz (Heyer, Quasthoff, & Wittig, 2012), S. 141 – 163
    ii. Themennotiz „Disambiguierung“ (Heyer, Quasthoff, & Wittig, 2012), S. 180ff
        1. Mehrdeutigkeiten begrenzen auf das passende Tauschelement

6. Offene Fragestellungen
    a. Wie kann man die Grammatik anpassen und umformen, jedoch basierend auf den Vorgaben des Autorenprofils des Eingabetextes und eventuellen Daten aus der Datenbank zu diesem Autorenprofil?
    b. Wie und welche Software zum „Maschine Learning“ kann eingesetzt werden, um die Eingabetexte im Hintergrund zu analysieren und den OneTIPP Algorithmus (automatisch) verbessernd anzupassen?

7. Qualitätskontrolle
    a. Entspricht der OneTIPP Ausgabetext dem Eingabe Autorenprofil?
    b. Wird die Eingabestimmung im Ausgabetext beibehalten?
    c. Speichere das Autorenprofil des OneTIPP Ausgabetextes für spätere Analysen
    d. Berechne Lesbarkeitsstatistiken vom Ausgabetext – dies sollte mit dem Eingabetext konform gehen („aus einem Bild.de Text soll kein wissenschaftlicher Text gemacht werden“ -> zumindest nicht in der ersten Version -> in den folgenden Versionen ist dies als aufwertendes Feature durchaus denkbar)
    e. Der „Pos-getaggte“ Ausgabesatz muss sich an den definierten Logiken für „Sinnhafte Kernelemente“ halten

Literaturverzeichnis

Heyer, G. (2012). *Text Mining: Wissensrohstoff Text.* Abgerufen am 16. November 2015 von Text Mining: Wissensrohstoff Text.: http://asv.informatik.uni-leipzig.de/uploads/document/file_link/462/TM13_Stimmungsanalyse.pdf

Heyer, G., Quasthoff, U., & Wittig, T. (2012). *Text Mining: Wissensrohstoff Text. Konzepte, Algorithmen, Ereignisse.* Witten: W3L-Verlag.

Roadmap Entwicklung OneTIPP

1. Lege ein Autorenprofil an
    a. Benutze dazu Hidden Markov Modelle
        i. Speicherung der Wahrscheinlichkeiten der auftretenden Wortketten als Autorenprofilfingerabdruck
        ii. Zeitaufwand: 5 h
    b. Führe eine Stimmungsanalyse durch (Heyer G. , 2012), S. 7ff
        i. Was sind bekannte und beliebte positive oder negative Schlagwörter des Autors? (Zeitaufwand: 4 h)
        ii. Zeitaufwand „Einfache Stimmungsanalyse“: 2 h
    c. (Speichere den POS-TAG Fingerabdruck des Autors – Welche Satzbaustrukturen werden gern verwendet?)
        i. Zeitaufwand: 1 h
    d. Berechnung der Wahrscheinlichkeiten eines Eingabesatzes – n-Gramme und Bestimmung der „Maximum Likehood Estimation“ (MLE) (Heyer, Quasthoff, & Wittig, 2012), S. 102
        i. Zeitaufwand: 8 h
    e. Berechne die Lesbarkeitsstatistiken des Eingabetextes
        i. Zeitaufwand: 1h