Gattungserkennung über 500 Jahre

Calvo Tello, José
jose.calvo@morethanbooks.eu
Universität Würzburg, Deutschland

Fragen

Wie gut lassen sich Gattungen und Untergattungen durch Maschinelles Lernen über eine längere Periode erkennen? Obwohl eine Reihe von Artikeln die Frage hauptsächlich für Englisch (Kessler, Numberg, und Schütze 1997; Petrenz und Webber 2011; Underwood 2014) und Deutsch (Hettinger et al. 2016) beantwortet hat, befasst sich wenig Forschung mit diesem Thema aus einer diachronischen Perspektive oder wird auf spanischen Texte angewendet (Henny-Krahmer 2018). Welche Gattungen sind leichter zu erkennen, welche komplizierter? Welche Algorithmen, Transformationen und Anzahl der lexikalischen Einheiten funktionieren am besten?

Datensatz: CORDE 1475-1975

Zur Beantwortung ob verschiedene Gattungen durch Maschinelles Lernen erkannt werden können, wurde das umfangreichste historische Korpus des Spanischen analysiert, CORDE. Dieses Korpus wurde von der Real Academia Española kompiliert (Rojo Sánchez 2010; Sánchez Sánchez und Domínguez Cintas 2007) und ist ein standard-Tool in der Hispanistik über das online Such-Interface (Kabatek und Pusch 2011). Für die Analyse wurden die Frequenzen der Tokens und die Metadaten jedes Texts an Forscher weitergegeben. Das Korpus beinhaltet ca. 300 Millionen Tokens (34.000 Texte) und die Texte sind mit expliziten Metadaten über Jahrhunderte, Länder und Gattungen markiert.

Die Daten der mittelalterlichen Sektion des Korpus präsentieren mehrere Probleme (Rodríguez Molina und Octavio de Toledo y Huerta 2017), wie beispielsweise ausgeprägte Unausgewogenheit der Anzahl der Texten im Vergleich zu anderen Jahrhunderten oder schwankende philologische Qualität. Deswegen wurden für diese Analyse nur die Texte der letzten 500 Jahre des Korpus selektiert, die länger als 100 Tokens sind. Somit beinhaltet das analysierte Korpus über 22.000 Texte (über 244 Millionen Tokens). Die Metadaten unterscheiden:

- Fachtexte in Themen (Jura, Geschichte, Geisteswissenschaften…)
- Gattungen und Untergattungen (lyrischer Vers, kurzer dramatischer Vers…)
- oder Medien (journalistische Texte, Briefe…).

Eine komplette Liste der Gattungen ist auf den Abbildungen zu finden.

Methoden der Evaluation

Die Klassifikation wurde binarisiert durchgeführt, d. h. jeder Text könnte zu jeder Gattung gehören oder nicht. Verschiedene Parameter wurden evaluiert:

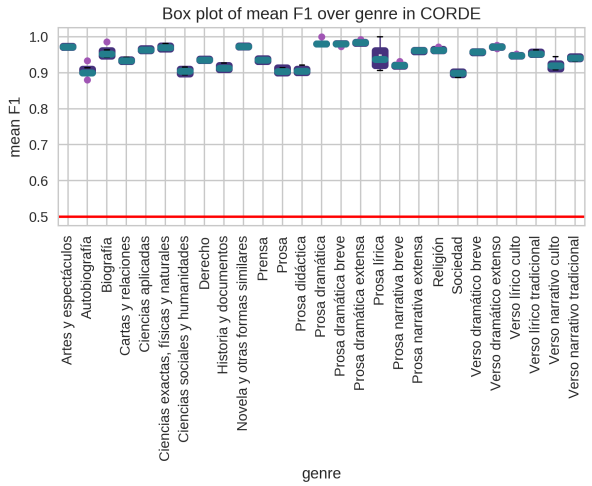
- Transformation der lexikalischen Information: relative Frequenz, binäre Frequenz, z-scores, TF-IDF, logarithmierte relative Frequenz
- Algorithmen: k-Nearest Neighbors, Random Forest, Logistic Regression und Support Vector Machine
- Anzahl der Tokens: 10, 50, 100, 500, 1.000, 2.000, 3.000, 4.000, 5.000 und 6.000

Das Korpus wurde für jede Gattung undersampled: die gleiche Anzahl an positiven wie an negativen Fällen wurden für jede Gattung gesamplet. Die Evaluation wurde mit Hilfe von Cross-Validation (10 folds) durchgeführt und der Mittelwert der F1 Scores berechnet. Der Code wird als Python Notebook über GitHub zugänglich sein.

Ergebnisse und Diskussion

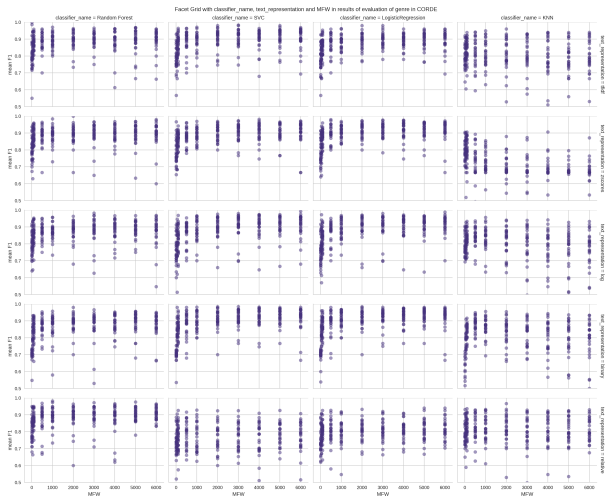
Die höchsten F1 Scores der Kombinationen von Parametern für jede Gattung lagen zwischen 0,9 und 1,0 mit einem Mittelwert der verschiedenen Gattungen von 0,96 (Standardabweichung von 0,03). Diese sehr hohen Ergebnisse ähneln sich denen von Underwood (2014), der an einem sehr großen Datensatz forschte. Die häufigsten Parameter bei den besten Ergebnissen waren Logistic Regression (16 Fälle von 27), binäre Häufigkeit (16, was nicht zu erwarten war) und 6.000 MFW (9).

Auf den nächsten Boxplots sind die 10 besten Kombinationen zu sehen. Jeder Punkt entspricht dem Mittelwert der F1 Scores der Cross-Validation der 10 besten Kombinationen von Parametern. Diese sind nach Gattung differenziert aufgelistet:



Folgende Gattungen wurden am besten erkannt: Theater (Vers und Prosa), Romane, lyrischer Vers, und Fachtexte über Naturwissenschaften und Kunst. Lyrische Prosa zeigt heftige Schwankungen, außerdem wurden die niedrigsten Ergebnisse von folgenden Gattungen erreicht: Autobiografie, narrativer Vers, Essay, lyrische Prosa, Prosa sowie Fachtexte über Gesellschaft, Geschichte und Geisteswissenschaften.

Ein interessanter Aspekt ist, welche die allgemeinen Tendenzen der Parameter und dessen Kombinationen sind. Dafür eignet sich ein Facet Grid Scatter Plot mit den Algorithmen als Spalten und den Transformationen als Reihen (einzelne Punkte entsprechen den Mittelwert der F1 Scores pro Gattung):



Hinsichtlich der Transformation (Reihen) zeigen die relative und die logarithmierte Häufigkeit niedrigere Ergebnisse als TF-IDF, z-scores und die binäre Häufigkeit. Bei den Algorithmen (Spalten) ist KNN merklich schlechter als die anderen drei. Zuletzt ist noch zu erkennen, dass die Qualität der Ergebnisse bis zu einer Anzahl von 2.000 Tokens zunimmt, und mit Schwankungen bis 6.000 stabil bleibt. Ein interessanter Aspekt ist die Tatsache, dass spezifische Kombinationen (SVC + TF-IDF, binäre + Logistic Regression, relative + Random Forest) von Vorteil im Vergleich zu anderen sind.

Bibliographie

Henny-Krahmer, Ulrike (2018): *"Exploration of Sentiments and Genre in Spanish American Novels."* In DH Conference. Mexico City: ADHO.

Hettinger, Lena / Reger, Isabella / Jannidis, Fotis / Hotho, Andreas (2016): *"Classification of Literary Subgenres."* In DHd Konferenz, 154–58. Leipzig: Universität Leipzig.

Kabatek, Johannes / Pusch, Claus D. (2011): *Spanische Sprachwissenschaft: eine Einführung*. Tübingen: Narr.

Kessler, Brett/ Numberg, Geoffrey / Schütze, Hinrich (1997): *"Automatic Detection of Text Genre."* In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics, 32–38. ACL '98. Stroudsburg, PA, USA: Association for Computational Linguistics.

Petrenz, Philipp / Webber, Bonnie (2011): *"Stable Classification of Text Genres."* Computational Linguistics 37 (2): 385–93.

Rodríguez Molina, Javier / Octavio de Toledo y Huerta, Álvaro Sebastián (2017): *"La imprescindible distinción entre texto y testimonio: el CORDE y los criterios de fiabilidad lingüística."* Scriptum digital: revista de corpus diacrònics i edició digital en llengües iberoromàniques, no. 6: 5–68.

Rojo Sánchez, Guillermo (2010): *"Sobre codificación y explotación de corpus textuales: Otra comparación del Corpus del español con el CORDE y el CREA."* Lingüística, no. 24: 11–50.

Sánchez Sánchez, Mercedes / Domínguez Cintas, Carlos (2007): *"El banco de datos de la RAE: CREA y CORDE."* Per Abbat: boletín filológico de actualización académica y didáctica, no. 2: 137–48.

Underwood, Ted (2014): *"Understanding Genre in a Collection of a Million Volumes*, Interim Report."