Informatiekunde, Rijksuniversiteit Groningen

26 november 2014 *Begeleider: Gosse Bouma*

**Onderzoeksvoorstel Information Retrieval**

*Classificatie van persoonlijkheidskenmerken*

**Rolf Kuipers *s2214415* - Mart Busger op Vollenbroek *s2174634***

Probleemstelling

Proberen de Big 5 (extravert, innovatief, stabiel, vriendelijk, zorgvuldig) te voorspellen op basis van sollicitatie- en motivatiebrieven, teneinde het automatisch en sneller evalueren van sollicitant.

Dataset

De gebruikte dataset zal bestaan uit 9000 sollicitatie- en motivatiebrieven inclusief persoonlijkheidstest van verschillende personen. De brieven bestaan uit geschreven tekst door een sollicitant. De persoonlijkheidstest bestaat uit schaalscores gegeven door de sollicitant zelf, over bepaalde persoonlijkheidskenmerken. Neem als voorbeeld een score van 4 op een schaal van 5 t.o.v. het kenmerk ‘innovatief’.

Deze complete dataset is bij de start van het onderzoek aangeleverd als een compleet .csv bestand met op iedere regel de motivatie en testwaarden van een enkel persoon.

Methode

Ons systeem zal in staat moeten zijn om zichzelf te trainen aan de hand van een groot aantal brieven en tests en vervolgens de persoonlijkheids-scores op de Big 5 te voorspellen voor een nieuwe set sollicitatie- en motivatiebrieven.

Om dit te kunnen doen zullen wij eerst de data op een gestructureerde manier moeten opslaan, zodat we hier later manipulaties op kunnen doen en experimenten mee kunnen uitvoeren.

De teksten zullen naar waarschijnlijkheid geclassificeerd worden aan de hand van de Naive Bayes methode, met daarnaast mogelijke andere classificeringsmethoden. Extra hulpmiddelen zullen naar alle waarschijnlijkheid een methode voor stemming (NL) zijn, stop-word filtering, gebruik van bi- en trigrammen en scores zoals de Jaccard Coefficient en de Levenshtein Edit Distance.

De persoonlijkheidstestscores zullen waarschijnlijk naar binaire schaal gebracht worden. Dit zal de volgende 5 classificaties opleveren (waarbij een score van 3 vervalt):

|  |  |
| --- | --- |
| Score 1/2 | Score 4/5 |
| Introvert | **Extravert** |
| Non-innovatief | **Innovatief** |
| Onvriendelijk | **Vriendelijk** |
| Instabiel | **Stabiel** |
| Onzorgvuldig | **Zorgvuldig** |

De gegeven score door een persoon zelf bepaald in welke categorie ze vallen. Zo zal de classificeerder brieven ook moeten kunnen plaatsen.

Evaluatie

Als het classificeringsprogramma op de juiste manier werkt, moet het in staat zijn om bijvoorbeeld willekeurig 90% van de data te nemen als training, en het programma te testen met de overige 10%. Als evaluatie moet het programma bijhouden in hoeveel van de gevallen de classificeerder een sollicitatie/motivatiebrief in de juiste persoonlijkheidscategorie plaats t.o.v. de daadwerkelijke door de persoon zelf gegeven score.

In hoeveel gevallen het programma dit goed classificeert zal dan de accuracy van de classificeerder zijn.

In de toekomst is een programma zoals dit, mocht het daadwerkelijk zeer goed werken, handig om op een snelle efficiënte manier sollicitatie- en motivatiebrieven te beoordelen naar persoonlijke kenmerken. Op deze manier kunnen zowel tijd als kosten bespaard worden.

Relevante literatuur

Om ons het onderwerp beter eigen te maken en een beter begrip te krijgen van text classifiers hebben we gezocht naar wetenschappelijke literatuur. We hebben in eerste instantie twee artikelen gevonden die dieper ingaan op text classifiers en de “Big five”. In het artikel van McCallum & Nigram wordt ingegaan op twee verschillende classifiers die beide de naam naive Bayes gebruiken, maar die op bepaalde punten grote verschillen vertonen. Judge et al. hebben onderzoek gedaan naar de “Big five” persoonlijkheidseigenschappen en het effect dat deze hebben op de mate van succes van de carrières van mensen.

Bibliografie

McCallum, A., & Nigam, K. (1998, July). A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization* (Vol. 752, pp. 41-48).

<http://staff.icar.cnr.it/manco/Teaching/2005/datamining/articoli/multinomial-aaaiws98.pdf>

Judge, T. A., Higgins, C. A., Thoresen, C. J., & Barrick, M. R. (1999). The big five personality traits, general mental ability, and career success across the life span. *Personnel psychology*, *52*(3), 621-652.

<http://people.tamu.edu/~mbarrick/Pubs/1991_Barrick_Mount.pdf>