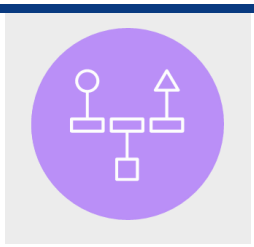
IBM Watson Natural Language Classifier



Automatische categorisatie	
Systeemvereisten:	Niet van toepassing
Ontwikkeld door:	IBM
Commerciële Licentie (SaaS)	Contactpersoon:	tom.ameloot@smals.be

Functionaliteiten

IBM Watson Natural Language Classifier (afgekort WNLC) laat toe om automatisch categorieën toe te kennen aan korte tekstfragmenten. Het product WNLC is gebaseerd op Deep Learning, een vorm van machine learning, waarmee geleerd wordt welke combinaties van woorden of volgorden van woorden kenmerkend zijn voor een bepaalde categorie. Een categorie is enkel een label; het label kan een zinnetje of een numerieke code zijn. Een belangrijke voorwaarde voor het gebruik van WNLC is dat er genoeg training data beschikbaar is. De training data bestaat uit een verzameling koppels van de vorm (korte tekst, categorie).

Er zijn vele mogelijke toepassingen denkbaar. Men zou bijvoorbeeld e-mails kunnen opdelen in verschillende categorieën, zoals algemene vragen, klachten, reclame, enz. Een ander voorbeeld, gegeven door IBM, is het bepalen of een uitspraak rond het weer specifiek gaat over weersomstandigheden (zoals sneeuw, mist) of temperatuur (warm, koud, gematigd); er zijn slechts twee categorieën in dit voorbeeld. In de context van de sociale zekerheid, zou een toepassing kunnen zijn om beschrijvingen van arbeidsongevallen te helpen omzetten naar de juiste juridische code, m.a.w., ondersteuning bieden bij het kiezen van een code uit een vaste lijst van mogelijkheden. Elke code is dan een aparte categorie.

WNLC geeft standaard meerdere suggesties terug van mogelijke categorieën, gerangschikt volgens afnemende matching score. Dit kan handig zijn bij interactie met eindgebruikers: als de matching score van de eerste categorie niet hoog genoeg is (wat afhangt van de toepassing), dan kan men bijvoorbeeld de beste drie categorieën tonen, waaruit de gebruiker vervolgens een keuze maakt.

Conclusies & Aanbevelingen

We hebben enkele testen uitgevoerd (zie volgende pagina), zowel op test-data van IBM zelf (rond het voorbeeld van het weer) als op eigen data. Onze conclusie is dat het product inderdaad goed is in datgene waarvoor het wordt aangeprezen, namelijk het toekennen van een categorie aan een korte tekst. Deze functionaliteit kan een mooie bouwsteen zijn in een groter project, of als toepassing op zich. Daarbij mag echter niet uit het oog verloren worden dat er inspanningen moeten geleverd worden om voldoende training data (voorbeeldzinnetjes) aan te bieden aan het systeem, wil men goede resultaten bekomen.

Testen & Resultaten

We hebben enkele testen uitgevoerd. In een eerste test hebben we de Engelstalige test-data van IBM gebruikt, die beschikbaar is in de handleiding van het product. Deze data bevat uitspraken rond het weer in twee categorieën: temperatuur (Engels: temperature) en weersomstandigheden (Engels: conditions). Deze data bevat onder andere de volgende voorbeelden:

Tekst	Categorie
How hot is it today?	temperature
How cold is it today?	temperature
What highs are we expecting?	temperature
What is today's expected humidity?	conditions
Will the blizzard hit us?	conditions
Is it drizzling?	conditions

Na het trainingsproces heeft het systeem inderdaad een goede feeling over welke woorden kenmerkend zijn voor welke categorieën. Bijvoorbeeld “hot” en “cold” duiden op temperatuur, en woorden zoals “drizzling” (motregen) duiden op weersomstandigheden. Hoewel men in dit voorbeeld in principe ook handmatig een programma zou kunnen schrijven om bepaalde woorden te detecteren, is het werkproces van WNLC makkelijker: (1) men verzamelt zinnetjes die representatief zijn voor de zinnetjes die het systeem moet kunnen afhandelen, en (2) men hangt aan elk zinnetje de gewenste categorie. Men moet dus geen data-analyse doen, of handmatig programmeren welke woorden (of combinaties daarvan) leiden tot bepaalde categorieën. Opmerking: Het is bij WNLC ook mogelijk om meerdere categorieën aan eenzelfde tekst toe te kennen.

In een tweede test hebben we met Franstalige zinnen gewerkt. We hadden namelijk beschikking over een Franstalig corpus van juridische codes van arbeidsongevallen, met bijhorende beschrijvingen. De bedoeling is om een code te selecteren op basis van de beschrijving. Als men de codes ziet als categorieën, dan kan men zo’n corpus beschouwen als training data voor WNLC. Onze test heeft uitgewezen dat ook hier het systeem leert welke woorden kenmerkend zijn voor bepaalde codes. Echter, deze test bevatte geen alternatieve verwoordingen voor eenzelfde code; er zal aanzienlijk meer training data moeten worden toegevoegd aan het corpus om deze case succesvol uit te werken voor eindgebruikers.

Een bijkomend advies, is dat men best overal een uniforme spelling zou moeten hanteren: als de training data correcte spelling gebruikt (met de juiste accent-tekens) dan moeten de eerstvolgende inputs (die moeten gecategoriseerd worden) ook de juiste spelling gebruiken; anders herkent het systeem de woorden niet.

Gebruiksvoorwaarden & Budget

De IBM Watson Natural Language Classifier is Software as a Service (SaaS), en draait in de IBM Bluemix cloud. Bij het moment van schrijven (maart 2017) is het mogelijk om één gratis categorie-model (Engels: “classifier”) te maken, met vier gratis training events per maand. Elk bijkomend model kost 20 US dollar per maand, en na de eerste 1000 categorisatie-oproepen kost een oproep 0.0035 US dollar.