De statistieken van Quantum Universe

Na veertien jaar en honderden artikelen over natuurkunde kijken we in dit stukje van Quantum Universe naar jullie, onze trouwe lezers. We bekijken met name de statistieken die in de loop der jaren over deze artikelen verzameld zijn. Met die blik hopen we zowel interessante feiten als natuurkundige patronen in de data te verduidelijken.

Aangezien het hedendaagse belang van machine learning groot is, maken we deels ook gebruik van die methoden om inzichten te zoeken. Tegenwoordig groeit de toepasbaarheid van machine learning net zo snel als de ondoorzichtigheid van de methoden die erin schuilgaan. Die methoden kunnen dus af en toe best geheimzinnig en zelfs oninteressant

lijken; zonder motivatie, redelijke basis of duidelijke denkkaders. Een deel van de bedoeling van dit artikel is dus om een eenvoudig voorbeeld te verstrekken waarin je het toepassen van eenvoudige machine learning-modellen als een leuke oefening kunt zien.

De artikelen op onze website gaan over allerlei soorten onderwerpen in de natuurkunde en de wiskunde, en we kunnen dus starten met hun thema's. Kunnen we die in een aantal categorieën verdelen? Er zijn natuurlijk al series (over onderwerpen als quantumfysica of zwarte gaten), en reeksen artikelen over verbonden onderwerpen, maar we kunnen proberen verder te zoeken naar een aantal grote thema's. Zo'n taak kan een natuurkundige relatief eenvoudig uitvoeren—je komt dan bijvoorbeeld op thema's als 'quantum', 'relativiteit', 'sterrenkunde', 'biofysica', 'wiskunde', of iets dergelijks—maar zo makkelijk is het voor een machine niet.

Clustering en unsupervised learning

In het algemene raamwerk van machine learning zijn er grofweg twee soorten problemen. Enerzijds kun je een model op een 'supervised' manier trainen, waarbij het doel op een precieze manier bepaald is en de prestatie van het model gemeten kan worden. Hier bestaat meestal een dataset waarvan een onderdeel gebruikt wordt om het model te trainen en de rest om het model te toetsen (in het Engels 'validation' genoemd). De meeste toepassingen van machine learning die je kent, vallen waarschijnlijk in deze categorie: neurale netwerken, transformers, Large Language Models (LLMs), enzovoort.

Als je geen kennis hebt van de patronen die je in de data zoekt, kun je anderzijds ook 'unsupervised learning' implementeren. Hier is er vooraf geen doel, maar in plaats daarvan probeert een algoritme te ontdekken wat er in de data schuilgaat. Je kunt bijvoorbeeld proberen clusters in je data te vinden, en hiervoor is het *k-means*-clusteringalgoritme het meest bekend.

Het idee van het k-means-algoritme is redelijk eenvoudig. Je wilt je data in k categorieën verdelen, waarbij elke categorie een bepaalde gemiddelde waarde heeft. Je begint dus met het raden van die k gemiddelde waarden. Elk datapunt deel je daarna in bij de categorie waarvan de gemiddelde waarde het dichtstbij ligt. Als dit klaar is, heb je volledige categorieën met een aantal datapunten, waarvan je een nieuwe gemiddelde waarde kunt berekenen, en daarna kun je de procedure herhalen totdat de categorieën niet of nauwelijks

meer verschuiven. Dit natuurlijke idee is succesvol toegepast op allerlei wetenschappelijke onderwerpen, vaak met prachtige resultaten.

Met verschillende technische hulpmiddelen kun je de titels van de artikelen als vectoren representeren, waarna het k-means-algoritme makkelijk uit te voeren is. De vectoren zijn zo opgebouwd dat elk woord dat tussen twee titels gedeeld wordt de overeenkomst vergroot. (Merk op dat zo'n benadering geen rekening houdt met de werkelijke *betekenissen* van de woorden, maar dat is ook bijna onmogelijk om op een bevredigende manier uit te voeren, en in ieder geval hebben we hier te weinig data om een taalkundig model te trainen.)

Als je dit proces toepast op de titels van QU-artikelen vind' je een verdeling in clusters, met de woorden die het meest bij elke categorie horen (het gebruikte programma begint te tellen bij groep 0):

Cluster 0:
complex, bliksem, chaos, complex natuur, dag, delen, dna

Cluster 1:
eerste, jaaroverzicht, donkere materie, echte, materie, holografie, lesmateriaal

Cluster 2:
nobelprijs, oerknal, delen, nieuwe, newton, deeltjes, schrodingers

Cluster 3:
bliksem, bouwstenen, deeltjes, complex natuur, actie, oerknal, gaten

Cluster 4:
actie, tussen, matrix, fouriertransformaties, terugkijken, getallen, quantumzwaartekracht

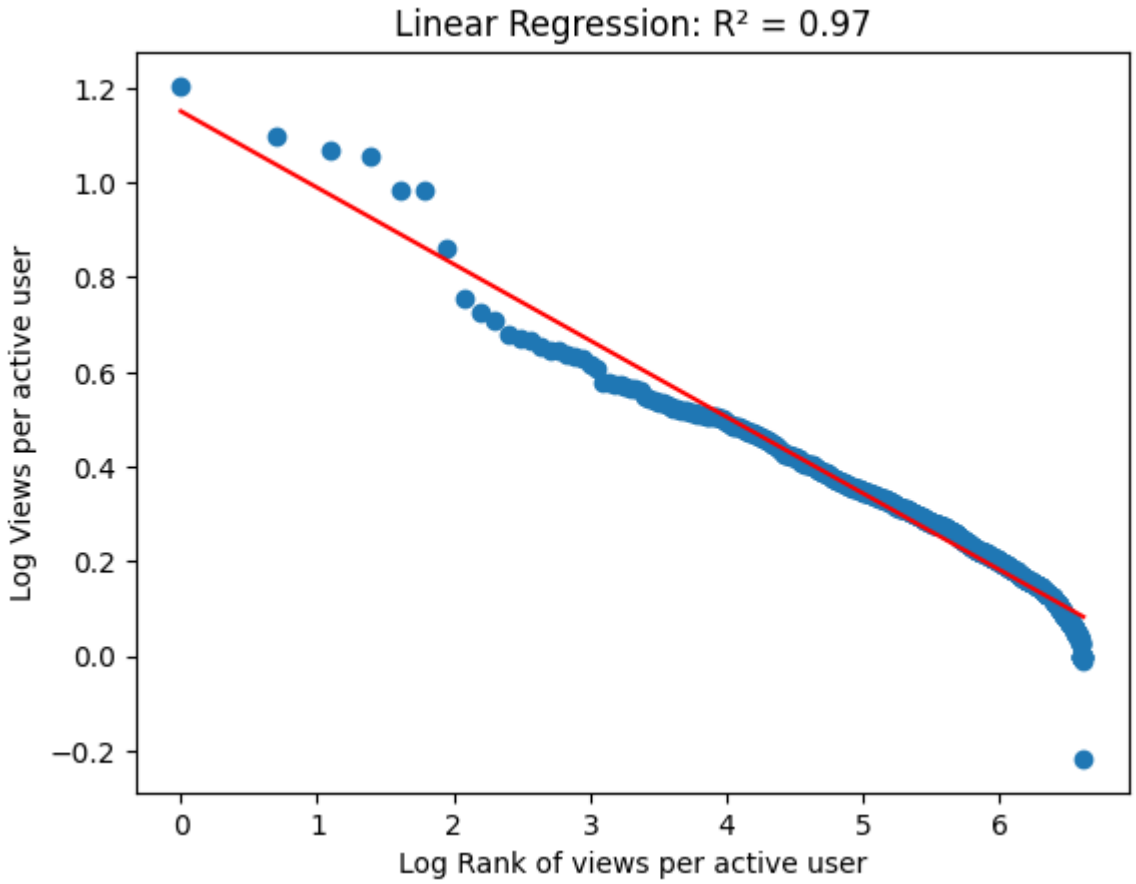
We zullen het straks hebben over hoe succesvol zo'n clustering is, maar er zijn al een paar patronen duidelijk als je de andere statistieken met de clusters vergelijkt. Als we bijvoorbeeld kijken naar de lengte van de artikelen in cluster 2 en de tijd die eraan besteed werd, zien we: allebei zijn ze veel korter, en inderdaad klopt dit, want 'nobelprijs' en 'oerknal' hebben meestal te maken met losse stukken bij speciale aankondigingen of gelegenheden.

cluster	Views	Average engagement time per active user	days_ago	word_count
0	197.863636	108.831235	2394.632231	1386.099174
1	172.223214	92.885542	1781.687500	1345.988839
2	235.000000	48.886477	2419.666667	620.333333
3	107.352941	88.612305	1785.235294	1309.470588
4	173.916667	119.916453	2451.777778	1821.250000

De natuurlijke wetenschappelijke vraag is dan hoe het nut van een bepaalde clustering gemeten kan worden. Er zijn wel technische hulpmiddelen zoals de zogeheten ‘elbow’ en ‘silhouette score’; die tonen in dit geval aan dat de clustering niet zo sterk is. Die conclusie kunnen we ook zónder diepgaande analyse al trekken, want een aantal woorden (‘bliksem’ of ‘actie’, bijvoorbeeld) verschijnt in verschillende categorieën. De belangrijkste oorzaak hiervan is dat alle titels binnen een relatief klein hoekje van de Nederlandse taal zitten, en de verschillen tussen titels met dezelfde woorden erin zijn voor zo’n eenvoudig algoritme moeilijk om te detecteren zonder een beter begrip van de taal en de natuurkunde.

Verdelingen en natuurkunde

Bijzonder interessant is de verdeling van hoe vaak elke actieve lezer een artikel bekijkt. Die kunnen we in een merkwaardig plotje zien:



In een eerder artikel had ik het over de wet van Zipf en het vreemde feit dat je in meerdere lijsten van frequenties ziet dat het tweede item twee keer minder vaak voorkomt dan het eerste, het derde drie keer minder, enzovoort. Iets algemener gezegd: ook als het patroon niet zo precies is, zie je vaak dat de frequentie waarmee een item voorkomt, afhangt van zijn rangorde volgens een machtwetrelatie (waarbij de macht in de wet van Zipf -1 is). Zo'n machtwetrelatie leidt tot een rechte lijn op een log-logplot, en je ziet hierboven dat het aantal maal dat een actieve lezer een link bekijkt heel goed zo'n rechte lijn volgt.

Er zijn veel theorieën voorgesteld om dit verschijnsel te verklaren. Een bekend voorbeeld heet 'preferential attachment'. Dit principe stelt dat artikelen die populair worden vaker gedeeld en vermeld worden, en dus een groter deel van de nieuwe kijkers aantrekken. Met een precieze formulering van zo'n stochastisch proces kun je de wet van Zipf afleiden. De macht die in ons geval goed bij de data past is blijkbaar veel lager in absolute waarde dan de door Zipf voorspelde -1 (we kunnen de macht eerder zien als de macht die bij de zogeheten paretoverdeling hoort). Deze macht heeft uiteindelijk met de sterkte van de 'preferential attachment' te maken, en het mechanisme hier heeft dus een andere sterkte, of misschien

zelfs een ander karakter, dan in het geval van de wet van Zipf.

In recent onderzoek werd voorgesteld dat de conclusie van een machtwetrelatie, hoewel het verleidelijk is die uit een plot te trekken, niet altijd geldt. Soms kan de juiste lijn iets subtieler zijn en de verdeling dus ingewikkelder dan een gewone machtwet. In ons plotje lijkt dat ook mogelijk te zijn, dus we moeten voorzichtig zijn met het al te snel trekken van conclusies.

Voorspelling van populariteit met supervised learning

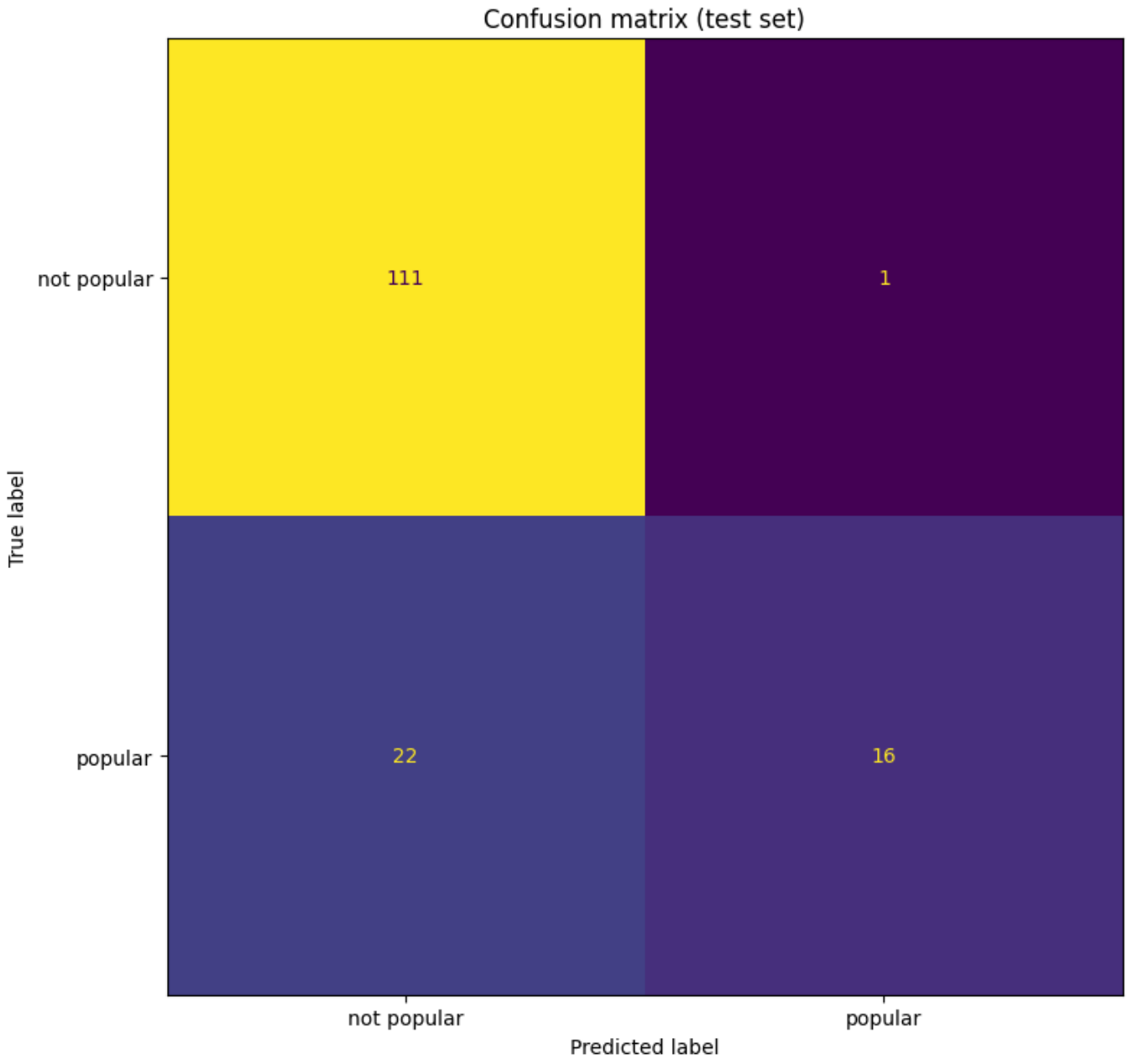
Een ingewikkelder en interessanter probleem is om de populariteit van een nieuw artikel te voorspellen op basis van eerdere artikelen. Aangezien het aantal lezers niet groot genoeg is om het precieze aantal views redelijk te kunnen voorspellen, definiëren we populariteit eenvoudiger, als “behorend tot de 25% meest bekeken artikelen”. Vervolgens moeten we een zogeheten classificatieprobleem oplossen, waarbij we proberen te leren van bepaalde kenmerken—de lengte van het artikel, de lengte van de titel, het onderwerp, onder andere—of een artikel in die zin wel succesvol zal worden of niet. Zodra we een supervised learning-probleem hebben, krijgen we daaruit ook direct een van de belangrijkste ingrediënten van machine learning: een loss-functie, die meet hoe dicht onze voorspellingen bij de echte resultaten liggen. In dit geval levert het model een kans voor elk artikel op, om top-25% of niet top-25% te zijn, en de loss-functie (die de ‘Cross-Entropy Score’ heet) houdt rekening met hoe zeker het model met zijn eigen conclusies is. Het punt is dus dat een model dat met veel vertrouwen foute voorspellingen maakt een ‘straf’ zou krijgen van de loss-functie.

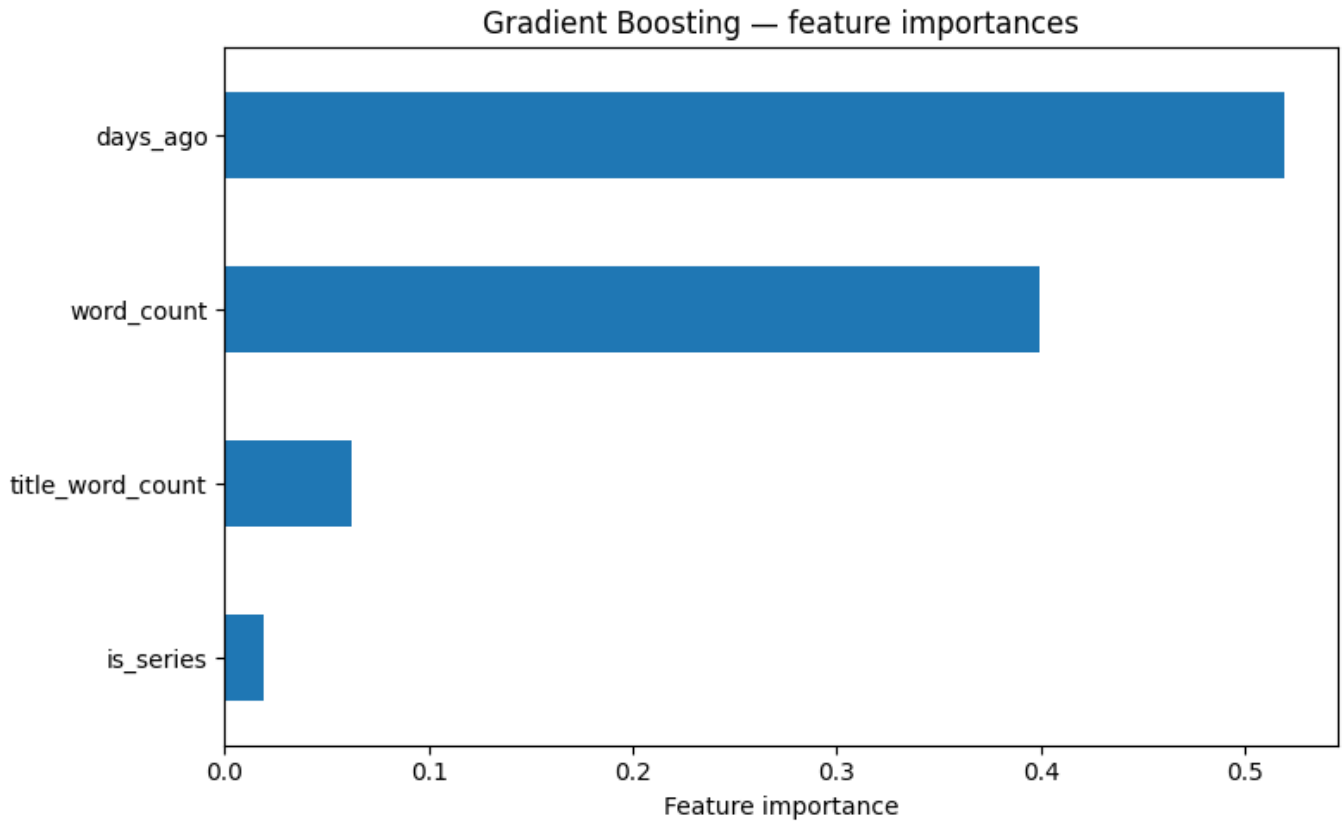
Voor het oplossen van het classificatieprobleem hebben we een nieuw model nodig, en in dit geval heb ik voor het zogeheten *gradient boosting*-model gekozen. Zonder technische details leg ik het idee kort uit: je wilt uiteindelijk beslissen of een artikel succesvol (1) of niet succesvol (0) zal worden. Je kunt hiervoor “decision trees” opbouwen—functies die bepaalde kenmerken met bepaalde gewichten meenemen. Het idee is dat je die gewichten kunt aanpassen om je classificatie zo goed mogelijk te maken. De innovatie van gradient boosting is dat je niet slechts één decision tree opbouwt, maar een reeks, waarbij je van de fouten van elke decision tree leert en de volgende verbetert door de minima van de loss-functie te zoeken.

Het presteren van zo’n model kunnen we nu meten met een maat die de ROC-AUC-score

heet. Die score is hier ongeveer 0,782 – een waarde die meestal als aanvaardbaar wordt gezien.

We kunnen op basis van de resultaten een paar interessante opmerkingen maken: ten eerste dat de lengte van de titel bijna niets te maken heeft met populariteit, en ten tweede dat het model heel terughoudend is om een artikel als populair te voorspellen, en dat ook niet zo goed doet.





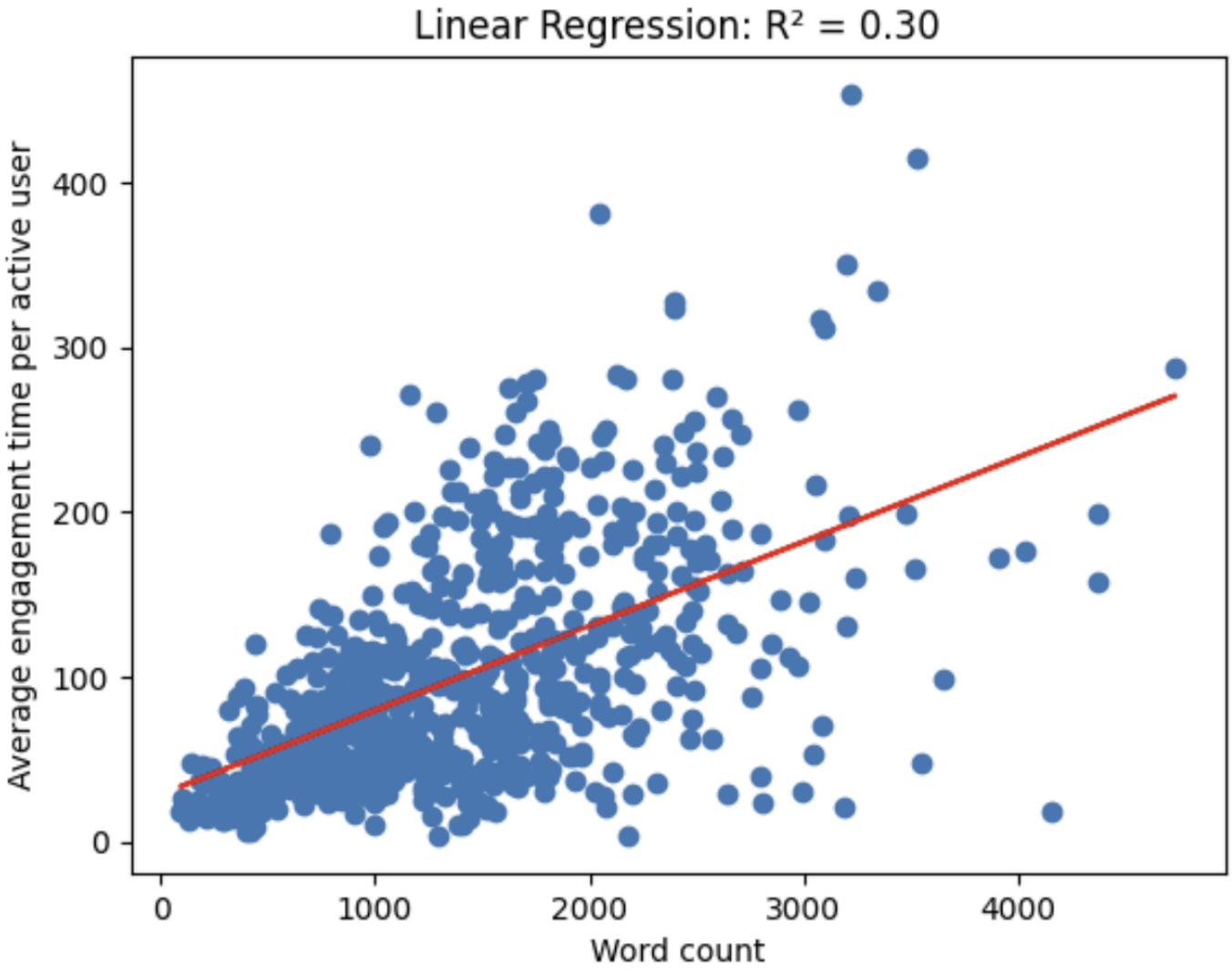
In ons model hebben we alleen betrekkelijk eenvoudige kenmerken gebruikt, maar als de clustering beter of ingewikkelder was geweest, hadden we die categorieën zeker ook kunnen gebruiken om een nog betere voorspelling te maken. Overigens heeft dit artikel zelf volgens het model slechts zo'n 17% kans om populair te worden! Ik blijf dus hopen dat machine learning-modellen nog steeds fouten kunnen maken…

Losse statistieken

Ten slotte verzamel ik wat losse statistieken die zowel voor de toevallige lezer als de toegewijde QU-liefhebber interessant kunnen zijn. We kijken eerst naar de artikelen in series—bijvoorbeeld 'quantummechanica' of 'relativiteit'—die in het algemeen veel populairder bleken te zijn dan de zelfstandige artikelen. De serieartikelen werden gemiddeld zo'n 246 keer bekeken, tegenover 163 voor de zelfstandige artikelen, en ze zijn ook meestal langer (met een gemiddelde lengte van 1768 woorden, tegenover 1282 voor zelfstandige artikelen).

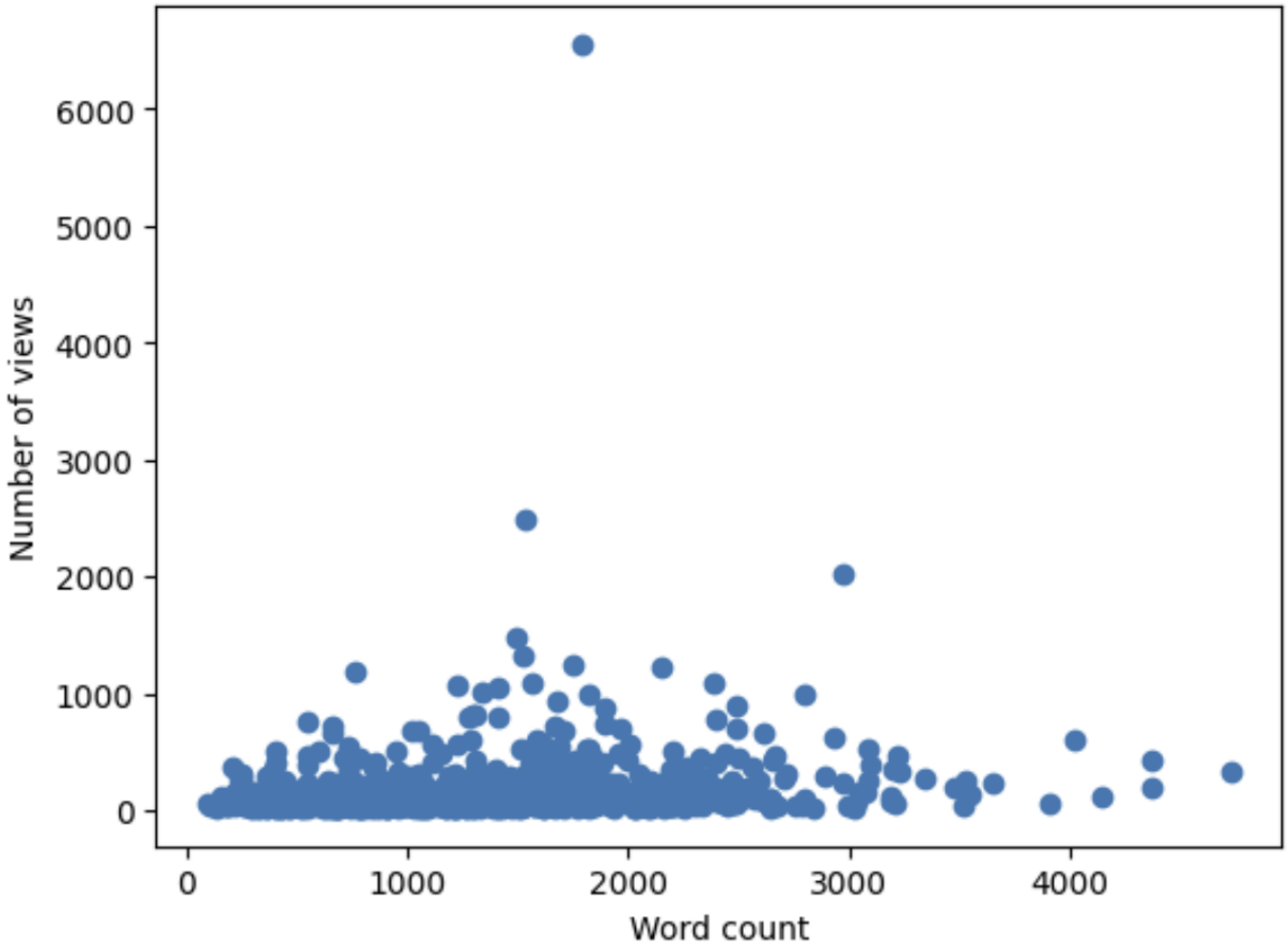
Je zou verwachten dat de lengte van een artikel te maken heeft met hoeveel tijd lezers eraan besteden. Op de Quantum Universe-website zien we dat jullie inderdaad meer tijd besteden

aan langere artikelen dan aan korte artikelen, wat volgens mij een goed teken is voor de publieke interesse in natuurkunde! (Let in de plot hieronder ook op de zogeheten R²-waarde, die meet hoe goed een rechte lijn bij de data past. In de gewone natuurkunde zou zo'n lage waarde van ~0,3 bijna nooit aanvaardbaar zijn, maar in een bijna sociologische context zoals we hier hebben is het wel relevant.)



Wat mij bijzonder opviel, is dat oudere artikelen zelfs in het afgelopen jaar heel veel aandacht kregen, soms zelfs meer dan nieuwe. We moeten vooral het artikel 'Hoe werkt een atoombom' vermelden, veruit het populairste op Quantum Universe met een ongelooflijke 6000 lezers in het afgelopen jaar, ondanks het feit dat het meer dan vijf jaar geleden gepubliceerd werd! Helaas zal de reden voor de grote interesse in dit artikel niet alleen de interessante natuurkunde zijn, maar ook het feit dat mensen in tijden met veel oorlogen meer op dit soort termen zoeken… Gelukkig zien we eenzelfde trend ook bij enkele andere

oude artikelen die nog steeds veel bekeken worden. Het aantal lezers in het afgelopen jaar blijkt uiteindelijk zels nauwelijks af te hangen van hoe lang geleden het artikel gepubliceerd werd, zoals je hier ziet:



Het is goed om te zien dat onze lezers de QU-artikelen in de afgelopen jaren al zo vaak bekeken hebben dat we er nu dit soort interessante statistieken uit kunnen halen. We verwachten nog heel lang door te gaan, en hopen onze dataset flink uit te breiden. Dus blijf klikken, lezen en vooral genieten!