Ejercicio Knime 1

Alberto Armijo Ruiz

Descripción del dataset utilizado

El conjunto de datos utilizado es un conjunto de datos que contiene noticias de la cadena de noticias BBC, el conjunto de datos se puede encontrar en el siguiente enlace: https://www.kaggle.com/shivamkushwaha/bbc-full-text-document-classification

Este conjunto de datos contiene 2225 documentos con 5 categorías diferentes; estas categorías son: *business, entertaiment, politics, sport* y *tech*. Para el ejercicio se ha utilizado los documentos de la categoría *tech*; esta categoría contiene 401 noticias sobre tecnología.

Preprocesamiento

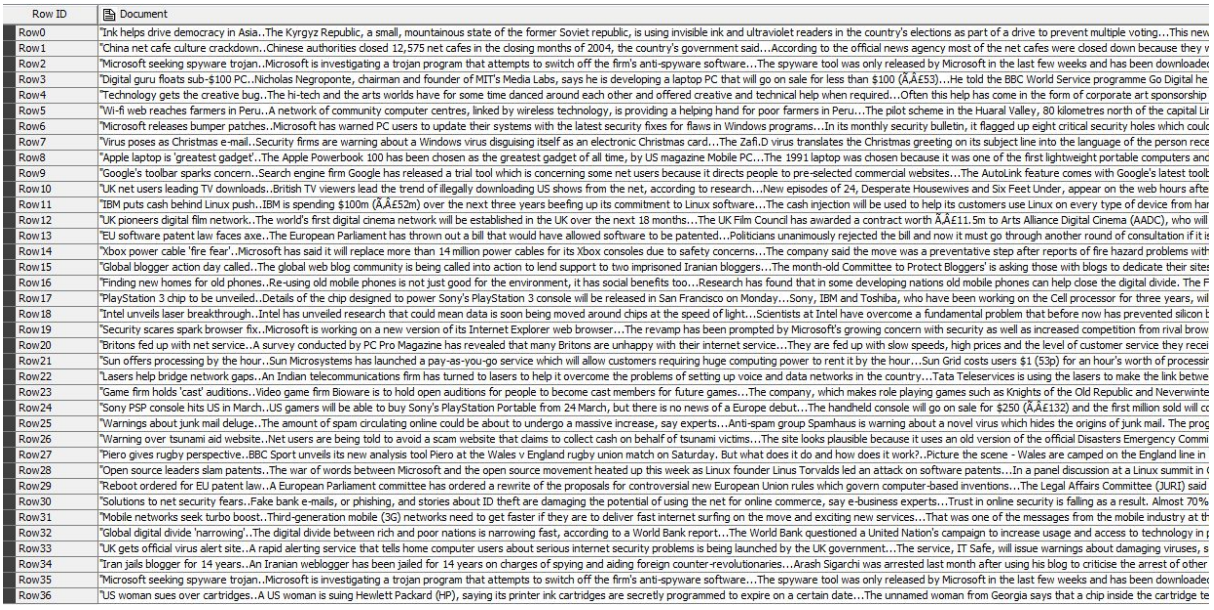
Para preprocesar los documentos lo primero que se ha hecho ha sido transformar los documentos, que primeramente en Knime se encuentra como texto, en tipo *document* y tras esto se puede borrar la columna de texto. Antes de esto, se ha tenido que transformar el conjunto de textos a un dataset; para ello se ha utilizado un script de python que lee los archivos y los introduce en un archivo csv.

Tras esto, se ha utilizado un POS Tagger para que analize cada uno de los documentos y obtenga la categoría de cada uno de estos (nombres, adjetivos, etc…). Una vez están etiquetados, se borran los signos de puntuación, palabras con menos de 3 caracteres, se le pasa un filtro para eliminar Stop Words (los artículos), se seleccionan todas las palabras que son nombres y se utiliza un Stemmer para obtener las raíces de las palabras en los documentos, de forma que trabajemos con estas y no con las distintas derivaciones de las palabras.

Después, se pasa a hacer una bolsa de palabras con todos los nombres que tiene el conjunto de documentos, la bolsa de palabras se utiliza para obtener un conjunto de términos por cada documento, que más tarde se utilizarán para calcular las frecuencias y encontrar los términos más importantes o más recurrentes en el dataset. Una vez creada la bolsa de palabras se calcula la frecuencia absoluta, relativa e IDF (inverse document frequency); de esta manera podemos saber la frecuencia (absoluta y relativa) que tiene cada término para el conjunto de datasets. Una vez obtenidos los datos, se calcula el valor tf-idf utilizando los valores de la frecuencia relativa e IDF. Por último, se filtran los resultados para quedarnos solamente con los términos que tienen alta frecuencia, de hecho nos quedaremos con los 10 términos con mayor frecuencia; y se muestran los resultados con un Cloud Tag.

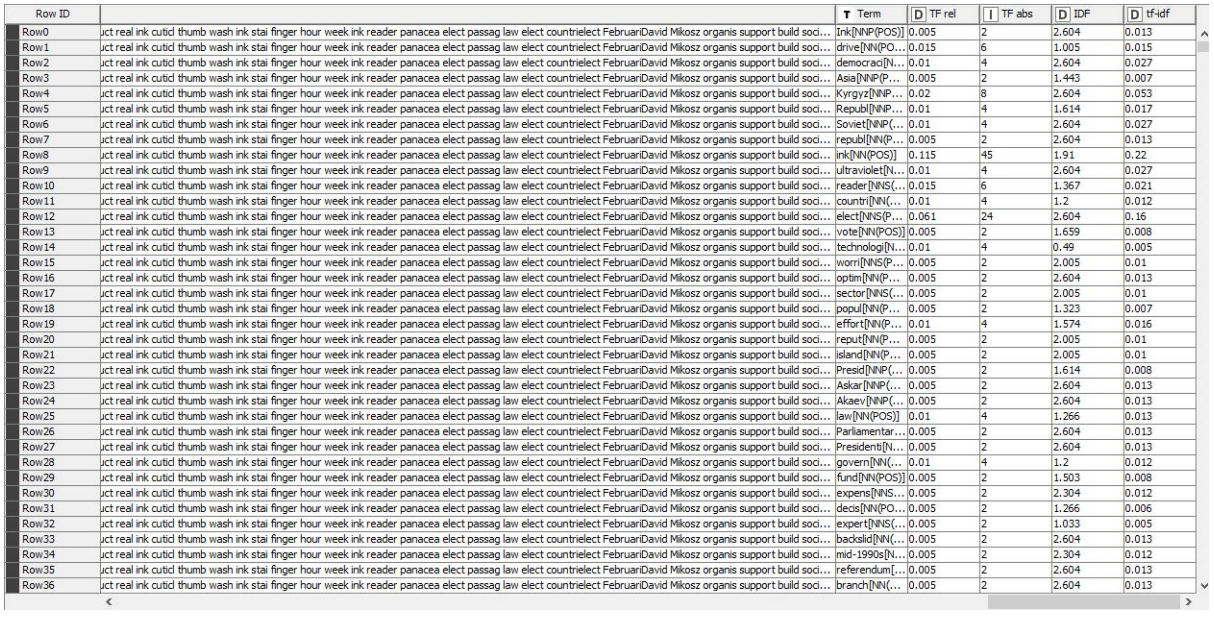
Para realizar el clustering, se han utilizado los datos antes de ser creada la bolsa de palabras, es decir, después de hacer stemming sobre ellos. Con ese conjunto de datos se filtran para obtener solamente las palabras más importantes y se realiza clustering sobre estos. Para el clustering se ha utilizado un nodo de clustering jerárquico, ya que a priori no podemos saber cuántos clusters hacer.

Tag Cloud y clustering



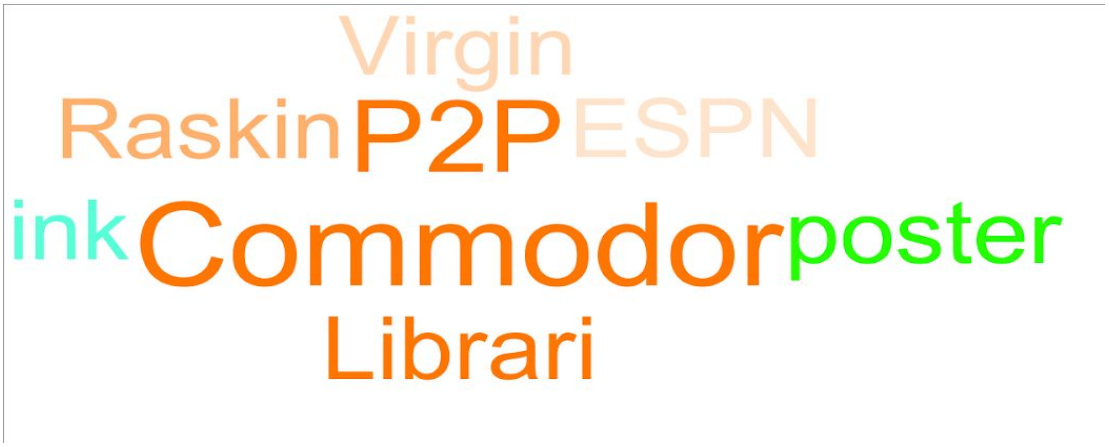
Row ID	Document
Row0	"Ink helps drive democracy in Asia..The Kyrgyz Republic, a small, mountainous state of the former Soviet republic, is using invisible ink and ultraviolet readers in the country's elections as part of a drive to prevent multiple voting...This new
Row1	"China net cafe culture crackdown..Chinese authorities closed 12,575 net cafes in the closing months of 2004, the country's government said...According to the official news agency most of the net cafes were closed down because they v
Row2	"Microsoft seeking spyware trojan..Microsoft is investigating a trojan program that attempts to switch off the firm's anti-spyware software...The spyware tool was only released by Microsoft in the last few weeks and has been downloaded
Row3	"Digital guru floats sub-$100 PC..Nicholas Negroponte, chairman and founder of MIT's Media Labs, says he is developing a laptop PC that will go on sale for less than $100 (Ã‚Â£53)...He told the BBC World Service programme Go Digital he
Row4	"Technology gets the creative bug..The hi-tech and the arts worlds have for some time danced around each other and offered creative and technical help when required...Often this help has come in the form of corporate art sponsorship
Row5	"Wi-fi web reaches farmers in Peru...A network of community computer centres, linked by wireless technology, is providing a helping hand for poor farmers in Peru...The pilot scheme in the Huaral Valley, 80 kilometres north of the capital Li
Row6	"Microsoft releases bumper patches..Microsoft has warned PC users to update their systems with the latest security fixes for flaws in Windows programs...In its monthly security bulletin, it flagged up eight critical security holes which could
Row7	"Virus poses as Christmas e-mail..Security firms are warning about a Windows virus disguising itself as an electronic Christmas card...The Zafi.D virus translates the Christmas greeting on its subject line into the language of the person rece
Row8	"Apple laptop is 'greatest gadget'..The Apple Powerbook 100 has been chosen as the greatest gadget of all time, by US magazine Mobile PC...The 1991 laptop was chosen because it was one of the first lightweight portable computers and
Row9	"Google's toolbar sparks concern..Search engine firm Google has released a trial tool which is concerning some net users because it directs people to pre-selected commercial websites...The AutoLink feature comes with Google's latest toolb
Row10	"UK net users leading TV downloads..British TV viewers lead the trend of illegally downloading US shows from the net, according to research...New episodes of 24, Desperate Housewives and Six Feet Under, appear on the web hours afte
Row11	"IBM puts cash behind Linux push..IBM is spending $100m (Ã‚Â£52m) over the next three years beefing up its commitment to Linux software...The cash injection will be used to help its customers use Linux on every type of device from har
Row12	"UK pioneers digital film network..The world's first digital cinema network will be established in the UK over the next 18 months...The UK Film Council has awarded a contract worth Ã‚Â£11.5m to Arts Alliance Digital Cinema (AADC), who will
Row13	"EU software patent law faces axe..The European Parliament has thrown out a bill that would have allowed software to be patented...Politicians unanimously rejected the bill and now it must go through another round of consultation if it is
Row14	"Xbox power cable 'fire fear'..Microsoft has said it will replace more than 14 million power cables for its Xbox consoles due to safety concerns...The company said the move was a preventative step after reports of fire hazard problems with
Row15	"Global blogger action day called..The global web blog community is being called into action to lend support to two imprisoned Iranian bloggers...The month-old Committee to Protect Bloggers' is asking those with blogs to dedicate their sites
Row16	"Finding new homes for old phones..Re-using old mobile phones is not just good for the environment, it has social benefits too...Research has found that in some developing nations old mobile phones can help close the digital divide. The F
Row17	"PlayStation 3 chip to be unveiled..Details of the chip designed to power Sony's PlayStation 3 console will be released in San Francisco on Monday...Sony, IBM and Toshiba, who have been working on the Cell processor for three years, wil
Row18	"Intel unveils laser breakthrough..Intel has unveiled research that could mean data is soon being moved around chips at the speed of light...Scientists at Intel have overcome a fundamental problem that before now has prevented silicon b
Row19	"Security scares spark browser fix..Microsoft is working on a new version of its Internet Explorer web browser...The revamp has been prompted by Microsoft's growing concern with security as well as increased competition from rival brow
Row20	"Britons fed up with net service..A survey conducted by PC Pro Magazine has revealed that many Britons are unhappy with their internet service...They are fed up with slow speeds, high prices and the level of customer service they recei
Row21	"Sun offers processing by the hour..Sun Microsystems has launched a pay-as-you-go service which will allow customers requiring huge computing power to rent it by the hour...Sun Grid costs users $1 (53p) for an hour's worth of processir
Row22	"Lasers help bridge network gaps..An Indian telecommunications firm has turned to lasers to help it overcome the problems of setting up voice and data networks in the country...Tata Teleservices is using the lasers to make the link betwe
Row23	"Game firm holds 'cast' auditions..Video game firm Bioware is to hold open auditions for people to become cast members for future games...The company, which makes role playing games such as Knights of the Old Republic and Neverwinte
Row24	"Sony PSP console hits US in March..US gamers will be able to buy Sony's PlayStation Portable from 24 March, but there is no news of a Europe debut...The handheld console will go on sale for $250 (Ã‚Â£132) and the first million sold will co
Row25	"Warnings about junk mail deluge..The amount of spam circulating online could be about to undergo a massive increase, say experts...Anti-spam group Spamhaus is warning about a novel virus which hides the origins of junk mail. The prog
Row26	"Warning over tsunami aid website..Net users are being told to avoid a scam website that claims to collect cash on behalf of tsunami victims...The site looks plausible because it uses an old version of the official Disasters Emergency Commi
Row27	"Piero gives rugby perspective..BBC Sport unveils its new analysis tool Piero at the Wales v England rugby union match on Saturday. But what does it do and how does it work?..Picture the scene - Wales are camped on the England line in
Row28	"Open source leaders slam patents..The war of words between Microsoft and the open source movement heated up this week as Linux founder Linus Torvalds led an attack on software patents...In a panel discussion at a Linux summit in C
Row29	"Reboot ordered for EU patent law..A European Parliament committee has ordered a rewrite of the proposals for controversial new European Union rules which govern computer-based inventions...The Legal Affairs Committee (JURI) said
Row30	"Solutions to net security fears..Fake bank e-mails, or phishing, and stories about ID theft are damaging the potential of using the net for online commerce, say e-business experts...Trust in online security is falling as a result. Almost 70%
Row31	"Mobile networks seek turbo boost..Third-generation mobile (3G) networks need to get faster if they are to deliver fast internet surfing on the move and exciting new services...That was one of the messages from the mobile industry at th
Row32	"Global digital divide 'narrowing'..The digital divide between rich and poor nations is narrowing fast, according to a World Bank report...The World Bank questioned a United Nation's campaign to increase usage and access to technology in p
Row33	"UK gets official virus alert site..A rapid alerting service that tells home computer users about serious internet security problems is being launched by the UK government...The service, IT Safe, will issue warnings about damaging viruses, s
Row34	"Iran jails blogger for 14 years..An Iranian weblogger has been jailed for 14 years on charges of spying and aiding foreign counter-revolutionaries...Arash Sigarchi was arrested last month after using his blog to criticise the arrest of other
Row35	"Microsoft seeking spyware trojan..Microsoft is investigating a trojan program that attempts to switch off the firm's anti-spyware software...The spyware tool was only released by Microsoft in the last few weeks and has been downloaded
Row36	"US woman sues over cartridges..A US woman is suing Hewlett Packard (HP), saying its printer ink cartridges are secretly programmed to expire on a certain date...The unnamed woman from Georgia says that a chip inside the cartridge te

Tabla inicial sin preprocesar



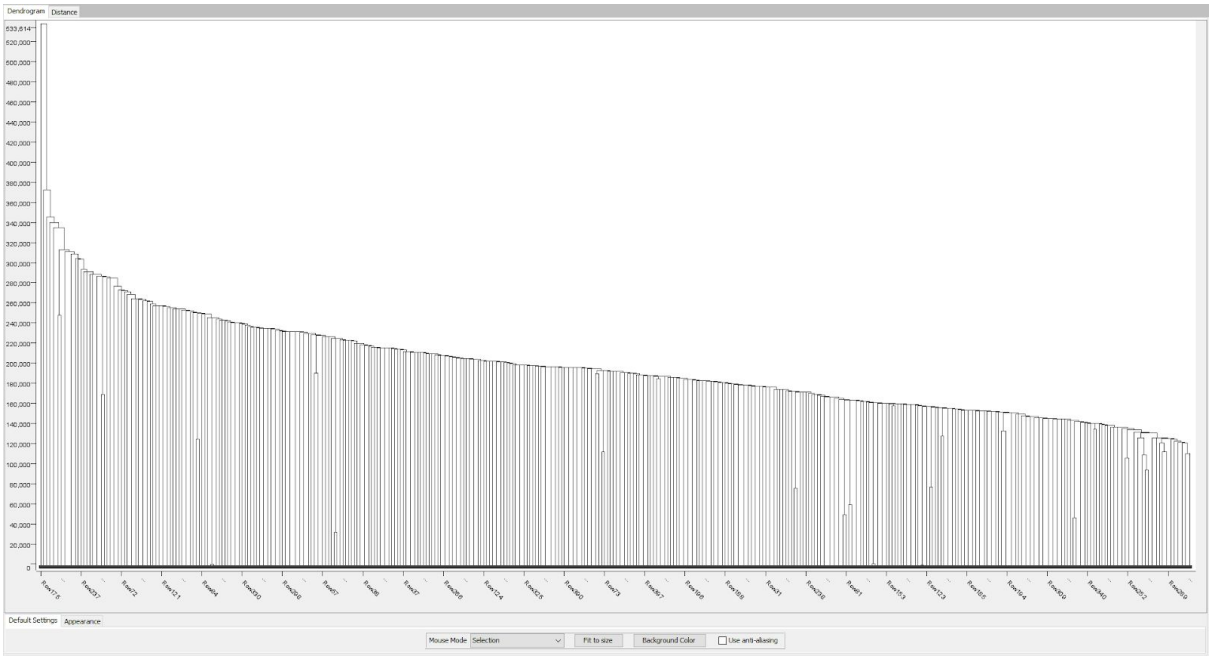
Row ID		Term	TF rel	TF abs	IDF	tf-idf
Row0	uct real ink cutid thumb wash ink stai finger hour week ink reader panacea elect passag law elect countrielect FebruariDavid Mikosz organis support build soci...	Ink[NNP(POS)]	0.005	2	2.604	0.013
Row1	uct real ink cutid thumb wash ink stai finger hour week ink reader panacea elect passag law elect countrielect FebruariDavid Mikosz organis support build soci...	drive[NN(PO...	0.015	6	1.005	0.015
Row2	uct real ink cutid thumb wash ink stai finger hour week ink reader panacea elect passag law elect countrielect FebruariDavid Mikosz organis support build soci...	democraci[N...	0.01	4	2.604	0.027
Row3	uct real ink cutid thumb wash ink stai finger hour week ink reader panacea elect passag law elect countrielect FebruariDavid Mikosz organis support build soci...	Asia[NNP(P...	0.005	2	1.443	0.007
Row4	uct real ink cutid thumb wash ink stai finger hour week ink reader panacea elect passag law elect countrielect FebruariDavid Mikosz organis support build soci...	Kyrgyz[NNP...	0.02	8	2.604	0.053
Row5	uct real ink cutid thumb wash ink stai finger hour week ink reader panacea elect passag law elect countrielect FebruariDavid Mikosz organis support build soci...	Republ[NNP...	0.01	4	1.614	0.017
Row6	uct real ink cutid thumb wash ink stai finger hour week ink reader panacea elect passag law elect countrielect FebruariDavid Mikosz organis support build soci...	Soviet[NNP(...	0.01	4	2.604	0.027
Row7	uct real ink cutid thumb wash ink stai finger hour week ink reader panacea elect passag law elect countrielect FebruariDavid Mikosz organis support build soci...	republ[NN(P...	0.005	2	2.604	0.013
Row8	uct real ink cutid thumb wash ink stai finger hour week ink reader panacea elect passag law elect countrielect FebruariDavid Mikosz organis support build soci...	ink[NN(POS)]	0.115	45	1.91	0.22
Row9	uct real ink cutid thumb wash ink stai finger hour week ink reader panacea elect passag law elect countrielect FebruariDavid Mikosz organis support build soci...	ultraviolet[N...	0.01	4	2.604	0.027
Row10	uct real ink cutid thumb wash ink stai finger hour week ink reader panacea elect passag law elect countrielect FebruariDavid Mikosz organis support build soci...	reader[NNS(...	0.015	6	1.367	0.021
Row11	uct real ink cutid thumb wash ink stai finger hour week ink reader panacea elect passag law elect countrielect FebruariDavid Mikosz organis support build soci...	countri[NN(...	0.01	4	1.2	0.012
Row12	uct real ink cutid thumb wash ink stai finger hour week ink reader panacea elect passag law elect countrielect FebruariDavid Mikosz organis support build soci...	elect[NNS(P...	0.061	24	2.604	0.16
Row13	uct real ink cutid thumb wash ink stai finger hour week ink reader panacea elect passag law elect countrielect FebruariDavid Mikosz organis support build soci...	vote[NN(POS)]	0.005	2	1.659	0.008
Row14	uct real ink cutid thumb wash ink stai finger hour week ink reader panacea elect passag law elect countrielect FebruariDavid Mikosz organis support build soci...	technologi[N...	0.01	4	0.49	0.005
Row15	uct real ink cutid thumb wash ink stai finger hour week ink reader panacea elect passag law elect countrielect FebruariDavid Mikosz organis support build soci...	worri[NNS(P...	0.005	2	2.005	0.01
Row16	uct real ink cutid thumb wash ink stai finger hour week ink reader panacea elect passag law elect countrielect FebruariDavid Mikosz organis support build soci...	optim[NN(P...	0.005	2	2.604	0.013
Row17	uct real ink cutid thumb wash ink stai finger hour week ink reader panacea elect passag law elect countrielect FebruariDavid Mikosz organis support build soci...	sector[NNS(...	0.005	2	2.005	0.01
Row18	uct real ink cutid thumb wash ink stai finger hour week ink reader panacea elect passag law elect countrielect FebruariDavid Mikosz organis support build soci...	popul[NN(P...	0.005	2	1.323	0.007
Row19	uct real ink cutid thumb wash ink stai finger hour week ink reader panacea elect passag law elect countrielect FebruariDavid Mikosz organis support build soci...	effort[NN(P...	0.01	4	1.574	0.016
Row20	uct real ink cutid thumb wash ink stai finger hour week ink reader panacea elect passag law elect countrielect FebruariDavid Mikosz organis support build soci...	reput[NN(P...	0.005	2	2.005	0.01
Row21	uct real ink cutid thumb wash ink stai finger hour week ink reader panacea elect passag law elect countrielect FebruariDavid Mikosz organis support build soci...	island[NN(P...	0.005	2	2.005	0.01
Row22	uct real ink cutid thumb wash ink stai finger hour week ink reader panacea elect passag law elect countrielect FebruariDavid Mikosz organis support build soci...	Presid[NNP(...	0.005	2	1.614	0.008
Row23	uct real ink cutid thumb wash ink stai finger hour week ink reader panacea elect passag law elect countrielect FebruariDavid Mikosz organis support build soci...	Askar[NNP(...	0.005	2	2.604	0.013
Row24	uct real ink cutid thumb wash ink stai finger hour week ink reader panacea elect passag law elect countrielect FebruariDavid Mikosz organis support build soci...	Akaev[NNP(...	0.005	2	2.604	0.013
Row25	uct real ink cutid thumb wash ink stai finger hour week ink reader panacea elect passag law elect countrielect FebruariDavid Mikosz organis support build soci...	law[NN(POS)]	0.01	4	1.266	0.013
Row26	uct real ink cutid thumb wash ink stai finger hour week ink reader panacea elect passag law elect countrielect FebruariDavid Mikosz organis support build soci...	Parliamentar...	0.005	2	2.604	0.013
Row27	uct real ink cutid thumb wash ink stai finger hour week ink reader panacea elect passag law elect countrielect FebruariDavid Mikosz organis support build soci...	Presidenti[N...	0.005	2	2.604	0.013
Row28	uct real ink cutid thumb wash ink stai finger hour week ink reader panacea elect passag law elect countrielect FebruariDavid Mikosz organis support build soci...	govern[NN(...	0.01	4	1.2	0.012
Row29	uct real ink cutid thumb wash ink stai finger hour week ink reader panacea elect passag law elect countrielect FebruariDavid Mikosz organis support build soci...	fund[NN(POS)]	0.005	2	1.503	0.008
Row30	uct real ink cutid thumb wash ink stai finger hour week ink reader panacea elect passag law elect countrielect FebruariDavid Mikosz organis support build soci...	expens[NNS...	0.005	2	2.304	0.012
Row31	uct real ink cutid thumb wash ink stai finger hour week ink reader panacea elect passag law elect countrielect FebruariDavid Mikosz organis support build soci...	decis[NN(PO...	0.005	2	1.266	0.006
Row32	uct real ink cutid thumb wash ink stai finger hour week ink reader panacea elect passag law elect countrielect FebruariDavid Mikosz organis support build soci...	expert[NNS(...	0.005	2	1.033	0.005
Row33	uct real ink cutid thumb wash ink stai finger hour week ink reader panacea elect passag law elect countrielect FebruariDavid Mikosz organis support build soci...	backslid[NN(...	0.005	2	2.604	0.013
Row34	uct real ink cutid thumb wash ink stai finger hour week ink reader panacea elect passag law elect countrielect FebruariDavid Mikosz organis support build soci...	mid-1990s[N...	0.005	2	2.304	0.012
Row35	uct real ink cutid thumb wash ink stai finger hour week ink reader panacea elect passag law elect countrielect FebruariDavid Mikosz organis support build soci...	referendum[...	0.005	2	2.604	0.013
Row36	uct real ink cutid thumb wash ink stai finger hour week ink reader panacea elect passag law elect countrielect FebruariDavid Mikosz organis support build soci...	branch[NN(...	0.005	2	2.604	0.013

Tabla inicial preprocesada



Resultados Tag Cloud con noticias de tecnología.

Por lo que se puede ver en los resultados del Tag Cloud, las palabras que más se repitieron para el conjunto de noticias sobre tecnología fueron nombres de empresas relacionadas con la tecnología, personas también relacionadas con la tecnología, nombres de tecnologías, etc...



Dendograma de los términos obtenidos

Por lo que se puede ver en el dendograma, no hay realmente clústers bien definidos, más bien existen clústers bastante pequeños de varias noticias relacionadas sobre temas parecidos.

Workflow Knime