

Práctica 1 Minería de Texto - Minería de Medios Sociales.

Curso: 2018/2019. Néstor Rodríguez Vico. DNI: 75573052C - nrv23@correo.ugr.es

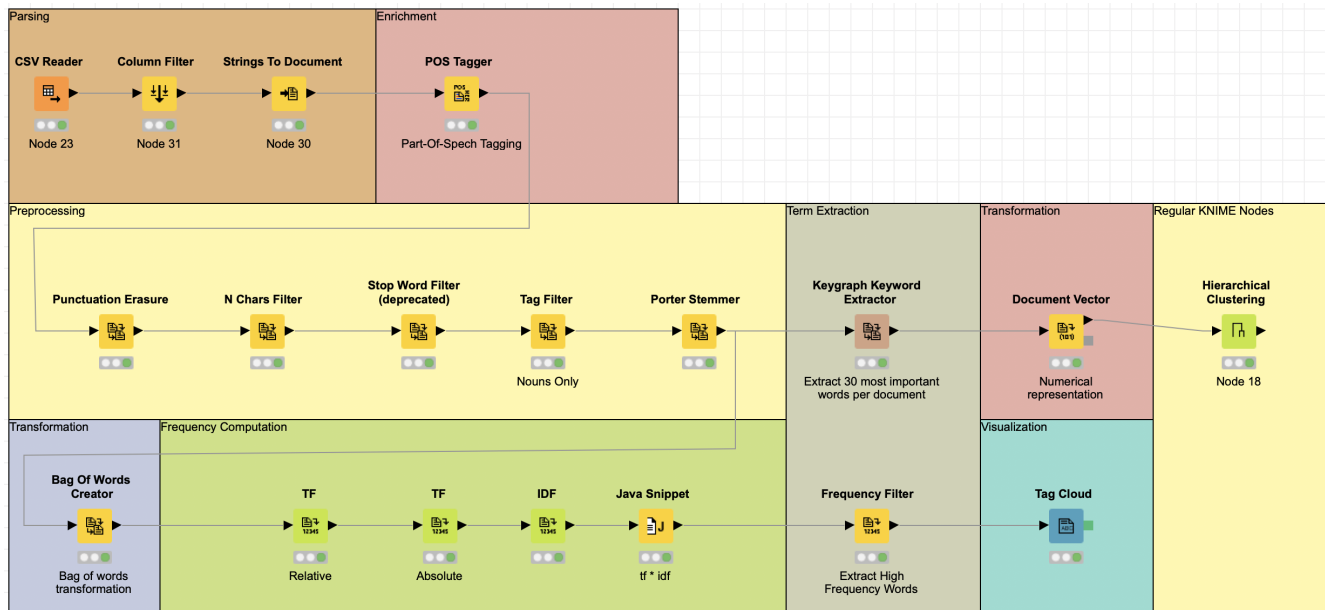
20 de mayo de 2019

1. Introducción.

Los datos usados para realizar esta práctica han sido obtenidos tras realizar la búsqueda *Documents classification* en *Google*. Dicha búsqueda lleva a muchas competiciones de *Kaggle*. Muchas de ellas son privadas (se necesita una invitación para poder descargar los dataset) pero en una de ellas (<https://www.kaggle.com/yufengdev/bbc-fulltext-and-category>) se puede descargar un conjunto de datos proveniente de la página de la BBC (<http://mlg.ucd.ie/datasets/bbc.html>). Esta colección consiste en 2225 documentos provenientes del sitio web de noticias de la BBC, correspondientes a historias de cinco áreas temáticas de 2004-2005. Dichas áreas son: negocios, entretenimiento, política, deporte, tecnología. Dicho conjunto de datos cuenta con 2225 documentos.

2. Workflow.

El *workflow* desarrollado es el siguiente:



Se ha cogido el *workflow* proporcionado y se han hecho los siguientes cambios:

- Se ha cambiado como se obtiene la colección de documentos. En este caso se ha empleado un lector de *CSV*. Una vez leídos los documentos, mandamos la información a un nodo que filtra las columnas para quedarnos solo con la columna que contiene la información del documento en sí.

- Una vez tenemos los datos limpios, convertimos el texto que representa el documento a una variable de tipo *Document*, que es el tipo usado por *KNIME* para representar un documento.
- En la sección de enriquecimiento he quitado el nodo *Abner Tagger* ya que es un preprocesamiento específico para el problema original propuesto y no aplica en el problema presentado en esta memoria.

El resto del *workflow* se mantiene de la misma forma:

- Eliminamos los signos de puntuación con el nodo *Punctuation Erasure*.
- Eliminamos las palabras con un tamaño menor que N , en nuestro caso 3, con el nodo *N Chars Filter*.
- Eliminamos las palabras vacías con el nodo *Stop Word Filter*. En nuestro caso, hemos usado la lista de *stopwords* proveída por *KNIME* para el idioma del problema, *inglés*.
- Eliminamos las palabras que tienen una etiqueta asociada, es decir, son palabras de una categoría concreta con el nodo *Tag Filter*. Por ejemplo, eliminamos los pronombres, ya que no nos aportan nada.
- Finalmente, aplicamos el algoritmo de *Porter* con el nodo *Porter Stemmer* para quedarnos con la raíz de las palabras.

Tras ejecutar el primer y segundo nodo ya tenemos el texto de los documentos cargados. Para ver el contenido de dicho nodo podemos pinchar en el segundo nodo y elegir la última opción del menú para mostrar los datos, la opción *Filtered table*. El resultado es el siguiente:

Row ID	text
Row0	tv future in the hands of viewers with home theatre systems plasma high-definition tvs and digital video recorders moving into the living room the way people watch tv will be radically different in f
Row1	worldcom boss left books alone former worldcom boss bernie ebbers who is accused of overseeing an \$11bn (£5.8bn) fraud never made accounting decisions a witness has told jurors david r
Row2	tigers wary of farrell gamble leicester say they will not be rushed into making a bid for andy farrell should the great britain rugby league captain decide to switch codes we and anybody else invo
Row3	yeading face newcastle in fa cup premiership side newcastle united face a trip to ryman premier league leaders yeading in the fa cup third round the game - arguably the highlight of the draw - is
Row4	ocean s twelve raids box office ocean s twelve the crime caper sequel starring george clooney brad pitt and julia roberts has gone straight to number one in the us box office chart it took \$40.8t
Row5	howard hits back at mongrel jibe michael howard has said a claim by peter hain that the tory leader is acting like an attack mongrel shows labour is rattled by the opposition in an upbeat speech
Row6	blair prepares to name poll date tony blair is likely to name 5 may as election day when parliament returns from its easter break the bbc s political editor has learned andrew marr says mr blair v
Row7	henman hopes ended in dubai third seed tim henman slumped to a straight sets defeat in his rain-interrupted dubai open quarter-final against ivan ljubicic the croatian eighth seed booked his pl
Row8	wilkinson fit to face edinburgh england captain jonny wilkinson will make his long-awaited return from injury against edinburgh on saturday wilkinson who has not played since injuring his bicep or
Row9	last star wars not for children the sixth and final star wars movie may not be suitable for young children film-maker george lucas has said he told us tv show 60 minutes that revenge of the sith w
Row10	berlin cheers for anti-nazi film a german movie about an anti-nazi resistance heroine has drawn loud applause at berlin film festival sophie scholl - the final days portrays the final days of the mer
Row11	virgin blue shares plummet 20% shares in australian budget airline virgin blue plunged 20% after it warned of a steep fall in full year profits virgin blue said profits after tax for the year to march w
Row12	crude oil prices back above \$50 cold weather across parts of the united states and much of europe has pushed us crude oil prices above \$50 a barrel for the first time in almost three months. fred
Row13	hague given up his pm ambition former conservative leader william Hague says he will not stand for the leadership again having given up his ambition to be prime minister mr Hague 43 told the
Row14	moya emotional after davis cup win carlos moya described Spain s davis cup victory as the highlight of his career after he beat andy roddick to end the usa s challenge in seville moya made up for
Row15	s korean credit card firm rescued south Korea s largest credit card firm has averted liquidation following a one trillion won (\$960m; £499m) bail-out lg card had been threatened with collapse bec
Row16	howard backs stem cell research michael howard has backed stem cell research saying it is important people are not frightened of the future the controversial issue was a feature of the recent us
Row17	connors boost for british tennis former world number one jimmy connors is planning a long-term relationship with the lawn tennis association to help unearth the next tim henman the american spe
Row18	japanese banking battle at an end japan s sumitomo mitsui financial has withdrawn its takeover offer for rival bank ufi holdings enabling the latter to merge with mitsubishi tokyo sumitomo bosses
Row19	games maker fights for survival one of Britain s largest independent game makers argonaut games has been put up for sale the london-based company behind the harry potter games has sacke
Row20	security warning over fbi virus the us federal bureau of investigation is warning that a computer virus is being spread via e-mails that purport to be from the fbi the e-mails show that they have c
Row21	halo 2 heralds traffic explosion the growing popularity of online gaming could spell problems for net service firms warns network monitoring company sandvine it issued the warning following anal
Row22	bates seals takeover ken bates has completed his takeover of leeds united the 73-year-old former chelsea chairman sealed the deal at 0227 GMT on Friday and has bought a 50% stake in the clu
Row23	cole faces lengthy injury lay-off aston villa s carlton cole could be out for six weeks with a knee injury the striker who is on a season-long loan from chelsea picked up the knock in an england un
Row24	mobile audio enters new dimension as mobile phones move closer to being a ubiquitous all-in-one media player audio is becoming ever more important but how good can that sound be from suc
Row25	moya fights back for indian title carlos moya became the first man to successfully defend the chennai open title by beating four-times finalist parashar srichaphan 3-6 6-4 7-6 (7/5) the spaniard
Row26	career honour for actor dicaprio actor leonardo dicaprio s exceptional career has been honoured at the santa barbara international film festival the star was presented with the award by martin s
Row27	mobile gig aims to rock 3g forget about going to a crowded bar to enjoy a gig by the latest darlings of the music press now you could also be at a live gig on your mobile via the latest third genera
Row28	terror suspects face house arrest uk citizens suspected of involvement in terrorism could face house arrest as part of a series of new measures outlined by the home secretary it comes after law lo
Row29	halloween writer debra hill dies screenwriter and producer debra hill best known for her work on the 70s horror classic halloween has died in los angeles aged 54 hill who had been suffering fro
Row30	royal couple watch nation s mood prince charles and camilla parker bowles are awaiting the nation s reaction after announcing they are to be married on 8 april mrs parker bowles will take the titl
Row31	firefox browser takes on microsoft microsoft s internet explorer has a serious rival in the long-awaited firefox 1.0 web browser which has just been released few people get excited when some ne
Row32	celebrities get their skates on former england footballer paul gascoigne will join eastenders actress scarlett johnson on bbc one s strictly ice dancing the one-off christmas special will also star tele
Row33	show over for mtv s the osbournes rock star ozzy osbourne has said his family will not make any more episodes of reality tv show the osbournes at the end of it it didn t like having cameras around
Row34	insurance bosses plead guilty another three us insurance executives have pleaded guilty to fraud charges stemming from an ongoing investigation into industry malpractice two executives from am

Tras ejecutar todos los nodos de la sección *Parsing*, de la sección *Enrichment* y de la sección *Pre-processing* tenemos un conjunto de datos ya preprocesado y limpio. Podemos ver como ha quedado dicho conjunto si hacemos click derecho en el nodo *Porter Stemmer* y seleccionamos la última opción del menú, la opción *Preprocessed documents*. El resultado es el siguiente:

Preprocessed documents. - 0/8 - Porter Stemmer	
File	Hilite Navigation View
Table: default - Rows: 2225 Spec - Columns: 2 Properties Flow Variables	
Row ID	Document
Row0	"hand viewer home theater plasma video record live people watch time panel consum electron technology impact passion trend program content viewer home network cabl satellit telecom compani servic provid devic talked-about technologi ce video record dvd parbox sh
Row1	"boss book boss berri-ebber fraud decisi juror myer comment defenc lawyer ebber phone compani prosecutor claim loss firm share myer fraud prosecutor defenc lawyer reid weingarten distanc client alleg cross exami myer ebber entri book weingarten m
Row2	"tiger gambi leicest bid rugbi leagu captain switch code stage tiger boss bbc radio leicest moment lot unknown situat gambi leicest oper knee week month leicest saracen list rugbi union club game union plai step leagu union centr progress post row rugbi leagu skill juri club t
Row3	"yead newcastl cup premiership newcastl trip rymen premier leagu leader yead cup round game highlight money-spinner yead round confer ewet citi dncast saturday travel traffard holder manchest home stake chelea plai host scunthorpe hineski brentford sundai leagu leader ep
Row4	"ocean raid box offic ocean crime caper sequel georg brad pitt julia box offic chart weekend ticket sale studio esqui sequel master crimin helst week treasur snipe blade trinit £84mtom hank comedi krangk ocean box offic triumph hank releas film ring trilog sequel beat predeces
Row5	"mongrel michael claim peter leader attack mongrel labour opposit speech parti spring confer labour campaign tactict tori home claim tactict bill debat bit leader common bbc radio stanc govern legial countri risk leader attack mongrel opposit opposit sake parti labour claim offic
Row6	"name poll date name elect dai parliam return break editor andrew marr qucen parliam week call name dai parti campaign street elect time matter minist spokswoman elect elect signal week westminist govern by kel legial parliam govern financ bill budget plan common busi sess
Row7	"head set public seed victori court level set progress rain public break seal dai seed robredo nicola kiedler germani weather umptir rain unipr control plai game score form desert"
Row8	"it captain jonni return injuri saturday bicep train newcastl fly-half saturday cup match bench newcastl director rugbi rob hope game autumn intern aggraw haematoma arm saracen captain jason robinson sale charli hodgson shirt intern africa injuri mustc month sidelin shoulder injuri v
Row9	"star war children star war movi children film-maker georg lucas minut reveng darkest seri cb program sundai lucas film rate parent scene film rate star war film guidanc rate except attack clone rate reveng prequel star war film chroncl transform anakin skywalker darth travel pi
Row10	"berlin cheer film german movi resist heroin applaus berlin film festi scholl dai portrai dai moviem scholl brother han condemn tyranit Adolf hitler director marc rothemund respons scholl idea film transcript interrog scholl trial archiv german secret public discover film rothemund r
Row11	"virgin share share budget arlin virn fall proffir virgin profit tax march demand novemb virgin chief bren godfreir virgin branson pressur rival jester passing forecast virgin fall quarter profit competit novemb half profit demand fuel cost virgin australia airlin market carrier qanta b
Row12	"oil price weather europ oil price barrel time month temperatur snowfal demand fuel stock valu dollar price mark time barrel oil york opec reason output peak barrel price averag brent rose trade europ north america temperatur dai dollar inflat price dollar oil chri fur market sta
Row13	"hagu ambt leader hagu stand leadership ambt minist hagu telegraph life polit hagu parti elect defeat rule return front bench paper hope north york famli wife hagu biograph pitt book newspap front rush stand leadership hagu determin role disappoint parti collin shadow ed
Row14	"carlo victori career roddick victori injuri beat roddick night dai energi tit australia outset peopl goal captain game opportunit time roddick dai posit victori perform nadal roddick singl dai malorcan player nadal coach patrick roddick rest team plai tennis clai home skill surfac quic
Row15	"korean credit card firm korea credit card firm liquid won card collas debt firm creditor parent rescu consortium creditor famili conglomer firm custom collaps shockwav countri economi firm creditor card deal week compani bankruptci delist compani move debt redempt com
Row16	"stem cell research michael stem cell research peopl issu featur elect georg bush tori leader embrac scienc victim parkinson motor neuron diseas duti offer million peopl stem cell stem cell master cell abil bodi tissu type scientist cell laborator form tissu kidnei brain tissu concern si
Row17	"conor boost tennis world comor relationship iwan tennis associ unearth tm dai la elit perform winner camp week attitud comor la arrang kid la chief john relationship coach player week hope relationship camp host junior player greg russekul arvid sie bunch kid comor que
Row18	"bank japan missui takeover offer bank ujf hold mitsubishi boss counterpart ujf decisi friday yen deal mitsubishi deal world bank asset ven exit fight bank histori ujf hold japan bank centr bid japan bank offer ujf compani ujf manag offer mitsubishi japan bank concern abil ujf defeat
Row19	"game maker fight surviv game maker game sale compani harri game employe cash crisi administ bbc new run cash low cash dai share trade stock exchang game game develop headquart north studio cambridg harri game flow cash compani software develop flow publish singl tro
Row20	"secur fi-viru bureau investig comput viru e-mails purport fbi e-mails address recipi website messag fbi fraud complaint center attach e-mail viru fbi messag recipi click attach answer questionnair attach viru infect comput agen viru comput user attach e-mails peopl recipi
Row21	"halo traffic explos popular game spel servic firm network monitor compani sandvin analysi traffic xbox game network launch dai halo novemb traffic explos sandvin servic provid network demand bandwidth titf halo network xbox gamer clan team comar surg demand bandwid
Row22	"bate seal takeover bate takeover leed chelea chairman deal gmt friday stake club club recognis leed club time lot club premiership fan bate stake guis compani sport fund plan bui leed elland road stadium arch train ground cours jon task stablis flow sort creditor bate tunnel matt
Row23	"cole injuri villa carbon cole week knee injuri striker loan chelea england under-21 match month carton action week challenge villa boss oper week oper chelea cole andl season rest shorlag striker return fit month andl emerg luke villa manag depart"
Row24	"audio dimens phone move media player audio devic bee jump disappear head demo home system phone british firm compani audio technology speaker firm offer stereo-widening technologi phone manag director monthest firm compani offer audio technologi bit headphon bit nice
Row25	"tight indian tit carlo chennai tit four-times finalist 6-4 spanish prize monel relief effort victim tsunami seed winner decol forc prize monel tournam tsunami victim differ live contribut pledg player prize monel tournam decemb disast live relief peopl tournam fee emergent relief fun
Row26	"career honour actor acori konardo career santa barbara film festi star award scores movi aviat lifetim achiev award film movi live life movi career honor film critter role basketbal dari titan gang achiev award california festi anniversary portray aviat actress russel film quest
Row27	"yig aim rock bar gig dari music gig genev rooster concert phone even venu gig phone phone oper technologi peopl video clip phone data network peopl phone download football music clip handest fan bandt pai pound ticket handest custom spokswoman b
Row28	"terror suspect hous arrest citizen involv terror hous arrest seri measur home secretari law lord detent terror suspect trial human right charl control terror hous arrest curfew law societ propos abus power deal deport detainee prison law terror attack effort director counti origin algeri
Row29	"halloween writer debora hill screenwrit produc debora hill 70s horror halloween lo angel hill suffer cancer co-wrote film lee curt psychopath john carpent record film time hill carpent york fog jersey hill career product assist rank assist director second-unit director carpent pionei
Row30	"coupl judge nate mood prind charl camilla parker nate reaction april parker tit duchess cornwal ceremon cast telegraph poll peopl two-thirds britton support coupl decisi parker consort charl king tit queen poll major briton monarchi gener queen hand throne grandison prind deat
Row31	"firefox browser internet explor firefox web browser peopl software program game music movi player releas version firefox drum amount fan software time version browser releas firefox novemb caus head program peopl software internet explor browser firefox mozilla foundat browse
Row32	"celebr skate footbal paul gasconen extend actress scarlett johnson bbc ic danc one-off christma star televis present carol smilli jessica celebri skater panel judg audicenc vote bbc star batt ic king queen present tress dalli host program heel night seri danc celebri practis ic danc ic
Row33	"osbourn rock star osbourn famili episod real osbourn dislin camera hous time sabbath singer report award come wife sharon osbourn famili life real osbourn seri watch episod dai sharon osbourn judg mentor talent x-factor loui walsh poll peopl rock career husband famili fo
Row34	"insur boss insur execut fraud charg investigi industri malpractice execut aig marsh mclemman investigi attorney spitzer plea rank execut marsh vice presidi bewla feloni crime prison marsh spokswoman bewla compani spitzer investigi insur industri compani bid price month marsh p

Si comparamos los resultados podemos ver las diferencias que hemos aplicado en los nodos. Por ejemplo, podemos ver como el documento representado en la línea identificada por *Row3* ha desaparecido el punto que hay. También podemos observar como han desaparecido las palabras con una longitud menor que 3. Aunque los cambios más notables son cómo hemos sido capaces de eliminar las *stopwords*, eliminar ciertas palabras que tienen una etiqueta asociada y el quedarnos con la raíz de las palabras. Estas tres técnicas nos permiten limpiar los documentos de palabras vacías de contenido (como pueden ser las *stopwords* o las palaras que tienen una etiqueta asociada y no nos interesan, como pueden ser los pronombres). Finalmente, nos hemos quedado con la raíz de las palabras, lo cual permite generalizar que texto es usado para entrenar nuestro modelo. Esta idea permite reconocer la palabra *comer* y la palabra *comes* como la misma, ya que su raíz es la misma.

Finalmente, usamos un nodo *Tag Cloud*, el cual genera una nube de etiquetas. Una nube de etiquetas es una representación de palabras que indica la importancia de las palabras de una forma visual, jugando con el tamaño y la transparencia del texto mostrado. En mi caso, las palabras más relevantes son *tottenham*, *librari* y *godzilla*, tal y como podemos ver en la nube de etiquetas generada por el *workflow*:

