Fase 2 – News

GDELT

Volendo studiare come "il mondo" parla degli incendi, abbiamo dovuto trovare un modo efficiente di recuperare il maggior numero di news provenienti da ogni parte del mondo. Abbiamo quindi fatto riferimento al dataset fornito dal progetto GDELT[1], che monitora le notizie mondiali dei telegiornali, dei quotidiani e del web da quasi ogni angolo di ogni paese, in oltre 100 lingue. Tale dataset fornisce il link dell'articolo e una serie di analisi ed informazioni già estrapolate dai testi degli articoli. Noi abbiamo utilizzato tale dataset solo per ottenere i link agli articoli che ci interessavano, per andare poi a scaricare noi stessi i testi e analizzarli a nostro piacere grazie agli strumenti appresi.

Il dataset è reso disponibile tramite la piattaforma Google BigQuery, attraverso la quale abbiamo fatto una prima interrogazione per selezionare solo le news provenienti da fonti di giornali, nel periodo 2015-2020[2], in cui fossero citate almeno una volta le parole "wildfire", "bushfire", "forest fire" o altre parole chiave legate agli incendi. Con solo tali filtri, però, le notizie cui avevamo accesso erano più di 1,5 milioni. Questo creava dei problemi sia per scaricare i dati (costi di BigQuery), che problemi di dimensionalità per i successivi passaggi di analisi. Abbiamo dunque deciso di selezionare solo alcune delle maggiori testate giornalistiche mondiali per la nostra analisi (vedere la tabella a lato).

Un primo criterio di decisione è stato quello di selezionare solo testate in inglese, per poter uniformare le analisi di text mining delle fasi successive. Abbiamo poi ricercato le principali testate per macro-aree geografiche: oltre ai sei continenti riconosciuti (Europa, Africa, Nord America, Sud America, Asia e Australia), abbiamo deciso di prendere in considerazione anche il Medio Oriente e la Russia, in quanto aree economicamente e politicamente rilevanti. Per ogni macro-area abbiamo dunque selezionato 6 testate giornalistiche (solo 3 per Australia e Russia, essendo singoli stati) che fossero presenti su GDELT. Sono state effettivamente scaricate più di 90mila news.

Macro-area	Stato	Testata giornalistica
AFRICA	(pan-africano)	allafrica.com
	Kenia	thestandard.com.hk
	Congo	africanews.com
	Zambia	iol.co.za
	Sud Africa	thesouthafrican.com
	(pan-africano)	africaleader.com
ASIA	Cina	asiatimes.com
	Giappone	japantimes.co.jp
	India	indiatimes.com
	Thailandia	bangkokpost.com
	Singapore/Malesia	asiaone.com
	Singapore	straitstimes.com
NORD AMERICA	Canada	cbc.ca
	Stati Uniti	latimes.com
	Stati Uniti	foxnews.com
	Stati Uniti	nbcnews.com
	Stati Uniti	nytimes.com
	Stati Uniti	cnn.com
SUD AMERICA	America latina	latinamericanpost.com
	America latina	latinpost.com
	Cile	santiagotimes.cl
	Cuba	plenglish.com
	Messico	santafenewmexican.com
	Brasile	riotimesonline.com
EUROPA	Inghilterra	dailymail.co.uk
	Inghilterra	economist.com
	Europa	euronews.com
	Inghilterra	reuters.com
	Europa	bbc.com
	Inghilterra	thetelegraph.com
MEDIO ORIENTE	Turchia	hurriyetdailynews.com
	Turchia	dailysabah.com
	Israele	jpost.com
	Israele	timesofisrael.com
	Iraq	aljazeera.com
	Arabia Saudita	arabnews.com
RUSSIA	Russia	themoscowtimes.com
	Russia	tass.com
	Russia	rt.com
AUSTRALIA	Australia	dailytelegraph.com.au
	Australia	theaustralian.com.au
	Australia	australiantimes.co.uk

[1] https://www.gdeltproject.org/data.html

[2] Le news recuperate partono dal 9 febbraio 2015, perché solo da tale data è iniziata l'implementazione del dataset di GDELT che abbiamo utilizzato per le nostre analisi, ovvero il GKG (Global Knowledge Graph) 2.0.

Le aree per cui sono stati recuperati più articoli sono il Nord America e l'Europa (per la quale la maggior parte delle news in inglese sono pubblicate da testate britanniche).[3] Particolarmente complicato è stato, invece, trovare testate giornalistiche in inglese per il Sud America, dove la maggioranza dell'informazione è in lingua spagnola o portoghese.

Scraping

Avendo accesso al link delle news, ma non al testo dell'articolo vero e proprio, è stato necessario fare *scraping* dai siti dei vari giornali per recuperare i testi. Per questo è stata utilizzata la libreria **Newspaper3k** che permette di automatizzare il *parsing* delle pagine html dei quotidiani, scaricando, tra altre cose, il titolo e il testo degli articoli. Non tutti i link erano, però, ancora attivi e non tutti gli articoli scaricabili gratuitamente, quindi vi è stata una naturale scrematura e sono rimaste quasi 53mila news.

Una volta scaricati i testi abbiamo dovuto effettuare un controllo manuale di ciò che era stato effettivamente recuperato, in quanto l'utilizzo di una libreria come newspaper per fare *scraping* presenta l'evidente vantaggio dell'applicabilità dello strumento a siti molto diversi ma, chiaramente, non sempre funziona come uno *scraper* costruito ad hoc sulla struttura html di un sito. È stato quindi necessario un controllo manuale per rimuovere eventuali messaggi di errore o richieste di sottoscrizione al giornale, confusi come testo dell'articolo in sé. Abbiamo anche eliminato i record che presentavano valori nulli per il testo dell'articolo, eliminato eventuali testi duplicati e controllato i titoli ripetuti.

Nonostante i giornali selezionati fossero testate internazionali che generalmente pubblicano notizie in lingua inglese, talvolta gli articoli recuperati erano in altre lingue. Grazie alla libreria **LangID** abbiamo dunque individuato la lingua in cui era scritto ogni articolo e abbiamo tenuto solo quelle effettivamente in inglese. Al termine di questi controlli ci sono rimaste più di 47mila news.

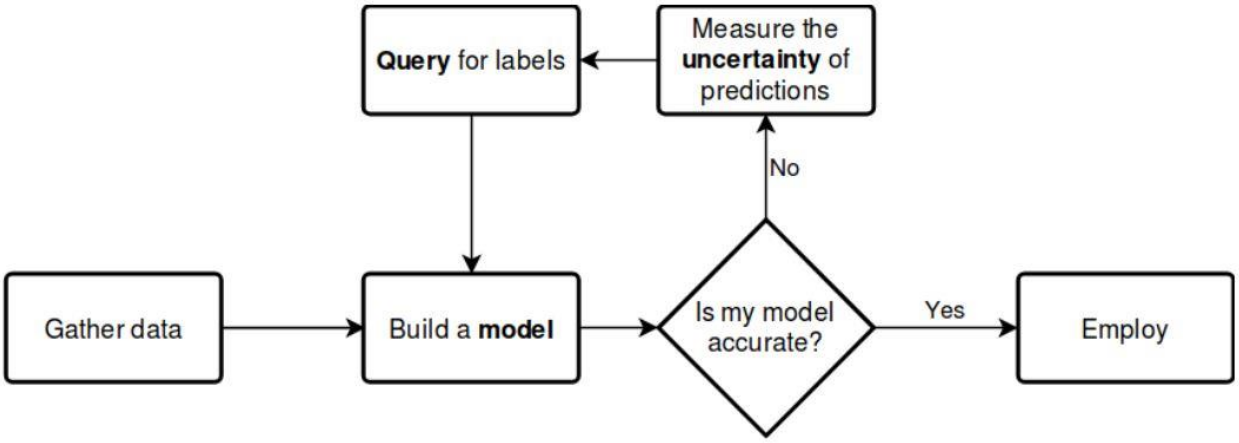
Interactive Learning

Un'ulteriore operazione di pulizia dei dati si è resa necessaria prima di procedere con l'analisi dei testi degli articoli: scaricando tutti gli articoli che contenevano almeno una volta una delle parole chiave, da noi definite, legate agli incendi, sono stati scaricati alcuni articoli che non parlavano di incendi forestali ma in cui la parola "wildfire" compariva soltanto, utilizzata in un altro contesto. Per selezionare solo le news che parlavano effettivamente di incendi forestali abbiamo dunque deciso di utilizzare l'interactive learning, una tecnica che permette di costruire un classificatore a partire da un numero minimo di news classificate a mano. Si parte ad allenare il classificatore su un piccolo numero di record etichettati a mano. Il classificatore restituisce le istanze sulla cui classificazione è più indeciso cosicché sia possibile fornirgli l'etichetta corretta. Il procedimento va avanti finché il supervisor non si stanca di classificare o non è sufficientemente soddisfatto dell'accuratezza del classificatore.

Per l'implementazione di tale tecnica abbiamo utilizzato la libreria **modAL**. Il task da svolgere si configura come un task di classificazione binaria ("0: non utile", "1: utile"). Il classificatore

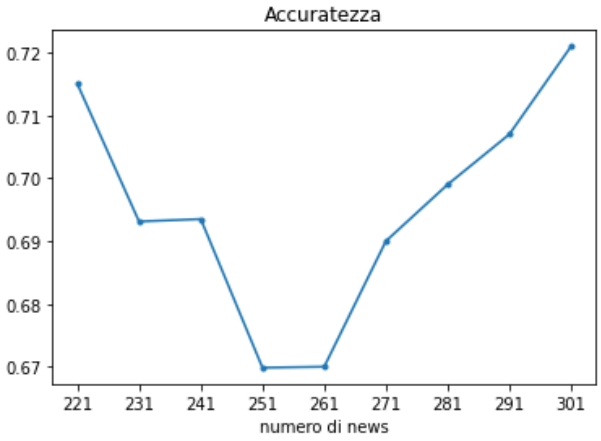
[3] Per quanto riguarda gli USA, abbiamo deciso di selezionare non solo testate giornalistiche di carta stampata ma anche reti televisive che trasmettono news, data la loro rilevanza per l'informazione, anche internazionale.

utilizzato è stato l'SVM e la strategia di proposta dei record da classificare a mano l'uncertainty sampling, in quanto strategia più efficace, come suggerito dalla letteratura[4].



Per poter dare del testo in input al classificatore è stato necessario fare una serie di operazioni di pulizia della stringa (rimuovendo caratteri speciali, punteggiatura e *stopwords*), abbiamo effettuato una sorta di *bolerplate detection* manuale rimuovendo i link grazie alle espressioni regolari, per poi effettuare la tokenizzazione e il calcolo del vettore TF-IDF.

Il primo dataset di train su cui abbiamo allenato il classificatore conteneva circa 220 news etichettate a mano. Ripetendo una serie di cicli di interactive learning siamo arrivate a classificare circa 300 news. Non avendo tutte le news etichettate, per valutare l'accuratezza è stato necessario effettuare la cross-validation sui dati di train.



Di lato riportiamo i risultati dell'accuratezza del classificatore in base al numero di news su cui si è allenato.

Ritenendo 72% di accuratezza un risultato discreto, abbiamo proceduto a classificare il resto delle news.

Per verificare l'accuratezza del classificatore abbiamo cercato tutti gli articoli in cui la parola "wildfire" compariva all'interno dell'articolo solo in espressioni tipiche della lingua inglese (ad esempio "spread like wildfire", che si traduce come "spandersi a macchia d'olio"). Tali articoli non parlano di incendi boschivi ma utilizzano solo quest'espressione pur parlando di tutt'altro. Abbiamo verificato che nella maggioranza dei casi il classificatore aveva riconosciuto, correttamente, come non utili queste news. Avendo un train sbilanciato sulla classe positiva (news utili), una piccola parte di queste news era stata classificata come news utile, creando alcuni falsi positivi. Abbiamo proceduto alla rimozione manuale di tali articoli, che consistevano comunque solo nel 4% dei casi positivi.

Le news che davvero parlavano di incendi boschivi erano, dunque, circa 40mila. Con queste abbiamo proceduto con le analisi sul testo vere e proprie.

[4] Moreo, A., Esuli, A., & Sebastiani, F. (2019). Building automated survey coders via interactive machine learning. *International Journal of Market Research, 4,* 408–429. DOI: 10.1177/1470785318824244

Prime analisi

In primo luogo abbiamo proceduto a studiare il numero di news per macro-area e per anno. Di seguito sono riportati i risultati:

Continent	
Africa	1430
Asia	4098
Australia	1059
Europe	15291
Middle_East	2285
North_America	15288
Russia	438
South_America	234

year	
2015	2770
2016	5760
2017	5358
2018	7359
2019	8455
2020	8805
2021	1616

Si può notare come le macro-aree più rappresentate siano Europa e Nord America mentre per la Russia e il Sud America sono state recuperate molte meno notizie. Per quanto riguarda le notizie per ogni anno si nota invece un costante aumento (bisogna escludere il 2021 per cui non sono presenti le news dell'intero anno). Questo è già di per sé un segno che sempre più si parla degli incendi forestali.

TAGME

Una domanda di ricerca cui eravamo interessati a dare risposta era come i vari paesi del mondo parlassero degli incendi. Per farlo abbiamo deciso di estrarre, per ogni macro-area geografica, le entità menzionate nei loro articoli. L'estrazione delle entità è stata effettuata tramite TAGME perché abbiamo notato che NER, di Spacy, riconosce meno entità (e queste sono soprattutto luoghi). TAGME invece, nonostante sia stato decisamente più dispendioso in termini di tempo, ha trovato un numero molto più elevato entità e molto differenziate tra loro (non solo luoghi). Dopo l'estrazione abbiamo calcolato l'occorrenza delle entità per ogni macro-area e considerato, per l'interpretazione, delle 50 più frequenti.

Un'altra domanda cui TAGME ci ha permesso di rispondere è di quali paesi si parli di più nel mondo e anche chi sia a parlarne. Infatti, dalla lista delle entità estratte per ogni macro-area abbiamo estratto solo i luoghi più frequenti per costruire una mappatura di "chi parla di chi".

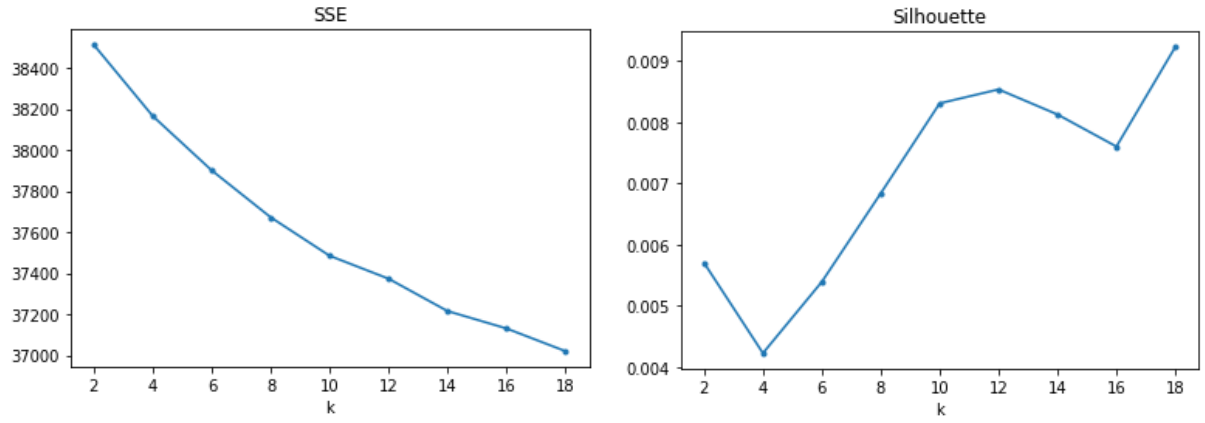
Topic modeling

Un'altra domanda di ricerca era come le news mondiali parlassero degli incendi. Abbiamo quindi cercato di estrarre i principali argomenti di cui si parla negli articoli grazie alla tecnica del topic modeling. Essa è però risultata poco adatta ai nostri dati in quanto i vari topic risultavano poco interpretabili e una buona parte degli articoli non veniva assegnata ad alcun topic.

Clustering

Abbiamo dunque proceduto con un'analisi di clustering, per avere una tecnica che ci fornisse un assegnamento sicuro di ogni articolo ad un cluster. L'algoritmo utilizzato è stato il K-means, cui abbiamo dato in input i vettori tf-idf relativi ad ogni articolo. Come metrica di distanza avremmo voluto usare la cosine-similarity (la più adatta per misurare la distanza tra documenti) ma non è implementata su SciKit Learn quindi abbiamo dovuto attenerci alla distanza euclidea prevista di default.

Per la model selection sono state provate nove configurazioni, con un numero di cluster da 2 a 18 (a salti di due). Di seguito sono riportati i grafici dell'SSE e della Silhouette.



Da tali grafici risulta che l'errore rimane sempre molto alto, nonostante la fisiologica diminuzione all'aumentare del numero di cluster, e la Silhouette risulta sempre vicinissima a 0. Questo potrebbe essere dovuto all'elevatissimo numero di feature presenti, dal momento che non è stato possibile fare feature selection (questa operazione è possibile solo in caso di dati etichettati, che noi non avevamo).

Abbiamo quindi optato per un modello a 5 cluster per una questione di interpretabilità dei cluster risultanti.

Per l'interpretazione dei cluster abbiamo estratto le parole più frequenti e le entità menzionate in ogni articolo, tramite la Named Entity Recognition fornita da Spacy. Di seguito i risultati:

	nome cluster	dimensione	top parole	top entità
0	Emergenza in California	1106	**PG&E** (Pacific Gas & Electric), **California**, fire, utility, **company**, state, wildfires, **bankruptcy**, people, customers, **billion**, **equipment**, safety, camp fire, winds, **power lines**, energy, **million**, **victims**, **killed**, Newsome, homes, weather	PG&E, **California**, San Francisco, Los Angeles, **Gavin Newsome**, 30 billion dollars, Northern California, californians, **California Public Utilities Commission**, Sacramento, Sonoma County, Butte County, **Bill Johnson**, San Diego, Santa Rosa, Southern California
1	Gestione dell'emergenza	23241	fire, people, **forest**, **wildfires**, state, **government**, California, home, water, **help**, national, **president**, air, city, **Trump**, oil, **work**, **smoke**, land, **need**, **firefighters**	**California**, Fort McMurray, canadian, Los Angeles, australian, **Australia**, Alberta, **Canada**, **Israel**, **Brazil**, **Indonesia**, american, summer, Democrats, Washington, **United States**, **amazon**, **China**, **Singapore**, Republicans, israeli, White House, indonesian, Congress
2	Cambiamento climatico	3464	**climate change**, global, **emissions**, world, people, **carbon dioxide**, **global warming**, weather, **temperatures**, energy, report, wildfires, fires, states, **heat**, countries, **greenhouse gas**, California, record, government, water, scientists, **Paris Agreement**, action, **extreme**, **environmental**, fire, state, national, need, Trump, **sea**, president, degrees, **forest**, air	**Paris Agreement**, **California**, UN, **Australia**, **United States**, **China**, summer, annual, **Europe**, australian, **Canada**, **India**, american, **Russia**, Washington, **Donald Trump**, **Alaska**, americans, **Arctic**, european, New York, Los Angeles
3	Emergenza in Australia	3880	fire, people, **Australia**, **bushfires**, Sydney, state, conditions, New South Wales, home, **firefighters**, **emergency**, **temperatures**, coast, **weather**, **burning**, **smoke**, service, australian, **residents**, **destroyed**, Victoria, **winds**, rural, water, rain, home, Morrison	**Australia**, New South Wales, australian, bush, **Scott Morrison**, south coast, south australia, summer, rage Australia, **koala**, East Gippsland, **Berejiklian**, Kangoroo Island, east coast, Sydney
4	Cronaca	8432	fire, **California**, **homes**, **people**, county, Los Angeles, **firefighters**, wildfire, **blaze**, **burned**, state, officials, **flames**, **winds**, **destroyed**, burning, **residents**, national, **evacuation**, forest, smoke, **weather**, department, lake, crews	**California**, Los Angeles, **Oregon**, Southern California, San Francisco, Santa Rosa, Ventura County, Northern California, Santa Barbara County, **Washington**, Sacramento, Sonoma County, Butte County, California department, Montecito, **Jerry Brown**, Santa Ana, **Colorado**, summer, **Carr Fire**