



Dipartimento di Ingegneria Informatica, Modellistica,
Elettronica e Sistemistica

Corso di Laurea Magistrale in Ingegneria Informatica

Tesi di Laurea

Sviluppo e confronto delle principali tecniche di
hashtag recommendation

Relatori:

Prof. Paolo Trunfio
Ing. Fabrizio Marozzo
Ing. Riccardo Cantini

Candidato:

Vincenzo Piperissa
Matricola 187747

Anno Accademico 2019/2020

Indice

Introduzione	1
1 Piattaforme di microblogging e Big Data	4
1.1 Microblogging	4
1.1.1 Twitter	6
1.1.2 Instagram	7
1.1.3 Vine	8
1.1.4 Tumblr	9
1.2 Big Data	9
1.2.1 Modello delle «3V»	11
1.3 Big Data Analytics	13
1.3.1 Modelli di programmazione	17
1.3.1.1 MapReduce	17
1.3.1.2 Functional programming	21
1.3.1.3 SQL-Like	23
1.3.1.4 Bulk Synchronous Parallel(BSP)	24
1.3.2 Database NoSQL	24
2 Tecnologie ed algoritmi utilizzati	31
2.1 Algoritmi basati su frequenza	32
2.1.1 TF-IDF	32
2.2 Clustering	34
2.2.1 DBSCAN	34
2.3 Modelli generativi	36
2.3.1 Latent Dirichlet Allocation (LDA)	37
2.4 Reti neurali	39
2.4.1 Recurrent Neural Network (RNN)	41
2.4.2 Long short-term memory (LSTM)	44

2.5	Attention	47
3	Stato dell'arte e dettagli implementativi	52
3.1	Hashtag Frequency-Inverse Hashtag Ubiquity (HF-IHU)	53
3.2	Tweet Embeddings e DBSCAN	57
3.3	Latent Dirichlet Allocation e campionamento di Gibbs	60
3.4	Global Co-Attention Bidirectional LSTM (GCA-BLSTM)	66
3.5	Topical Co-Attention (TCAN)	69
4	Analisi dei risultati	75
4.1	Analisi del dataset	75
4.2	Recall	77
4.3	Confronti tra le tecniche	83
4.3.1	Hashtag recommendation	83
4.3.2	Polarization	84
	Bibliografia	90

Elenco delle figure

1.1	Modello delle 3V	11
1.2	Ciclo di vita dei Big Data	16
1.3	Modelli di programmazione	17
1.4	Architettura framework Apache Hadoop	18
1.5	MapReduce: Word Count	20
1.6	Architettura Spark	22
1.7	Hive: Word Count in HQL	23
1.8	Teorema CAP	26
1.9	Esempio Key-Value datastore	28
1.10	Esempio Document datastore	28
1.11	Esempio Graph datastore	29
1.12	Esempio Column-Family datastore	30
2.1	Formula TF-IDF	33
2.2	DBSCAN: Pseudocodice	35
2.3	DBSCAN Data points	36
2.4	Modello LDA	37
2.5	Formula LDA	38
2.6	Variazione della distribuzione θ al variare di α	39
2.7	Schema neurone artificiale	40
2.8	Rete neurale feed-forward	41
2.9	Recurrent Neural Network	42
2.10	Applicazione ripetuta della funzione sigmoide	43
2.11	Regola di aggiornamento del gradiente	43
2.12	Long short-term memory (LSTM)	45
2.13	Forget gate	46
2.14	Input gate	46
2.15	Output gate	46
2.16	Modello Seq2seq	47

2.17	Architettura Encoder-Decoder con meccanismo di Attention	48
2.18	Matice di allineamento	49
2.19	Self Attention	50
3.1	Algoritmo di ottimizzazione per la creazione delle liste THFM e HFM	55
3.2	Smooth Inverse Frequency (SIF) proposta nel paper di Arora [33]	59
3.3	Pseudocoide campionamento di Gibbs	62
3.4	Rappresentazione delle probabilità tramite barre	63
3.5	Andamento perplexity al variare del numero di topic	65
3.6	GCA-BLSTM Model Summary	68
3.7	Rappresentazione grafica del modello GCA-BLSTM	69
3.8	Rappresentazione ad alto livello del modello TCAN	70
3.9	TCAN Model Summary	73
3.10	Rappresentazione grafica del modello TCAN	74
4.1	Esempio struttura Tweet	75
4.2	Divisione tweet all'interno del dataset	76
4.3	HF-IHU: calcolo della recall variando il coefficiente di espansione n	78
4.4	DBSCAN: calcolo della recall variando il coefficiente di espansione n	79
4.5	LDA-GIBBS: calcolo della recall variando il coefficiente di espansione n	80
4.6	GCA-BLSTM: calcolo della recall variando il coefficiente di espansione n	81
4.7	TCAN: calcolo della recall variando il coefficiente di espansione n	82
4.8	Confronto Recall tenendo fisso il coefficiente di espansione pari a 0	83
4.9	Confronto Weighted Recall variando il coefficiente di espansione n	84
4.10	Confronto Polarizzazione	86

Introduzione

Nell'ultimo decennio lo sviluppo tecnologico ha coinvolto persone di ogni fascia di età e la tecnologia è entrata a far parte delle nostre vite introducendo verso di essa una sorta di dipendenza; basta darsi uno sguardo attorno per poter osservare le teste chinate e gli occhi fissi e puntati sullo schermo dello smartphone a leggere le ultime news di cronaca, di gossip o gli aggiornamenti delle persone seguite sui social network. Ed è proprio in questo contesto che hanno preso sempre più piede i microblog, riscontrando un grande successo e assumendo un ruolo di una certa importanza nella vita quotidiana di ognuno di noi.

Le piattaforme di Microblogging sono un insieme di messaggi istantanei di lunghezza limitata che consentono all'utente iscritto al servizio, chiamato blogger, di condividere eventi, idee o sensazioni con altre persone della stessa piattaforma. Queste piattaforme sono diventate popolarissime, e fra tutte queste Twitter ne è l'esempio lampante: con oltre 300 milioni di utenti attivi e bilioni di visite mensili, Twitter ricopre una posizione fra le più alte nella classifiche delle piattaforme di Microblogging. Gli utenti iscritti a questo servizio possono postare messaggi e condividere contenuti che prendono il nome di tweet la cui lunghezza è limitata a 280 caratteri. Ogni tweet può contenere la presenza di un segno, il cancelletto '#', che precede una parola cliccabile in cui non vi sono spazi o segni di punteggiatura: questo termine prende il nome di hashtag. Un utente può inserire un hashtag nel tweet con lo scopo di specificare un topic per quel tweet oppure categorizzarlo in modo tale che altri bloggers lo trovino cercando altri tweet basati sullo stesso argomento. Questa funzionalità, insieme a tante altre, moltiplicata per il bacino di utenza della piattaforma genera una mole notevole di dati che possono essere analizzati utilizzando tecniche specifiche per trarre informazioni utili: tale quantità assume il termine tecnico di Big Data. L'analisi dei queste informazioni mediante algoritmi di NLP (Natural

Language Processing) ha portato allo sviluppo di task differenti fra cui assume una certa rilevanza la raccomandazione di hashtag (Hashtag recommendation).

Il seguente lavoro di tesi ha come scopo l'analisi, l'implementazione ed il confronto delle maggiori tecniche dello stato dell'arte per l'hashtag recommendation; verranno illustrate nel dettaglio le metodologie e le tecnologie usate, le scelte tecniche adoperate nell'implementazione, gli step necessari che vanno dal pre-processing del dataset alle considerazioni sui risultati ottenuti.

L'elaborato si articola in quattro capitoli. Il primo capitolo riguarderà i social network e i big data: in particolare verranno descritte le principali piattaforme di microblogging, come queste influenzano la vita quotidiana di ognuno di noi, successivamente si farà un analisi sui dati generati da tali servizi, sulle caratteristiche e su gli usi che li coinvolgono.

Il secondo capitolo mostrerà nel dettaglio le tecniche dello stato dell'arte usate per l'hashtag recommendation.

Il terzo capitolo, a sfondo prettamente tecnico, conterrà la spiegazione di ogni metodologia accompagnata da pseudocodice, gli approcci utilizzati e le scelte tecniche. Le tecniche comprendono varie metodologie, a partire da algoritmi basati su frequenza che peccano di semantica fino ad arrivare a meccanismi più recenti come l'Attention, che rappresenta lo stato dell'arte nell'ambito della comprensione del testo.

Infine, il quarto e ultimo capitolo, esporrà l'analisi dei risultati ottenuti e i confronti fra le varie tecniche. I risultati si baseranno principalmente su una metrica chiamata Recall e su una percentuale di polarizzazione, che indicano rispettivamente la capacità dell'algoritmo di predire correttamente gli hashtag dato un tweet in input e la capacità di individuare le preferenze politiche espresse in relazione ad un insieme fissato di hashtag notoriamente a favore di una data fazione o un candidato. Le metodologie implementate sono state valutate su un dataset di tweet a sfondo politico, nello specifico i tweet fanno riferimento alle elezioni presidenziali del 2016 negli Stati Uniti caratterizzate dalla rivalità tra Hillary Clinton e Donald Trump. I tweet analizzati risultano inoltre geo localizzati all'interno di dieci stati precisi, in particolare Wisconsin, Colorado, Florida, Iowa, Michigan, New Hampshire, North Carolina, Ohio, Pennsylvania e Virginia, caratterizzati da un'elevata incertezza circa il candidato presidenziale

vincitore. Lo scopo delle metodologie sarà, quindi, quello di restituire l'insieme dei top-k hashtag più rilevanti per il tweet in input, dai quali è possibile ricavare il grado di polarizzazione verso i due candidati, ovvero le preferenze espresse tramite gli hashtag usati.

Bibliografia

- [1] Apache spark - introduction. https://www.tutorialspoint.com/apache_spark/apache_spark_introduction.htm. Accessed: 2020-06-14.
- [2] Apache spark: la piattaforma per elaborare i big data. <https://lorenzogovoni.com/apache-spark-lo-strumento-per-elaborare-i-big-data/>. Accessed: 2020-06-14.
- [3] Attention? attention! <https://lilianweng.github.io/lil-log/2018/06/24/attention-attention.html>. Accessed: 2020-06-19.
- [4] Attention in neural networks. <https://towardsdatascience.com/attention-in-neural-networks-e66920838742>. Accessed: 2020-06-19.
- [5] Attn: Illustrated attention. <https://towardsdatascience.com/attn-illustrated-attention-5ec4ad276ee3>. Accessed: 2020-06-19.
- [6] A beginner's guide to lstms and recurrent neural networks. <https://pathmind.com/wiki/lstm>. Accessed: 2020-06-19.
- [7] Big data. https://it.wikipedia.org/wiki/Big_data. Accessed: 2020-06-06.
- [8] Big data analytics. https://it.wikipedia.org/wiki/Big_data_analytics. Accessed: 2020-06-07.
- [9] Cosa sono i big data: esempi concreti della vita quotidiana. <https://www.cloudtalk.it/big-data-esempi/>. Accessed: 2020-06-07.
- [10] Dbscan clustering in ml | density based clustering. <https://www.geeksforgeeks.org/dbSCAN-clustering-in-ML-density-based-clustering/>. Accessed: 2020-06-19.

- [11] Hadoop: il sistema più utilizzato per gestire i big data. <https://lorenzogovoni.com/hadoop-per-gestire-i-big-data/>. Accessed: 2020-06-09.
- [12] I tre tipi di analisi dei big data: descrittive, predittive e prescrittive. <https://www.nextre.it/tipi-di-analisi-dei-big-data/>. Accessed: 2020-06-07.
- [13] Il microblogging: una nuova forma di blog. <https://www.ionos.it/digitalguide/online-marketing/social-media/il-microblogging-blogging-compatto-in-maniera-semplice/>. Accessed: 2020-06-04.
- [14] Illustrated guide to lstm's and gru's: A step by step explanation. <https://towardsdatascience.com/illustrated-guide-to-lstms-and-grus-a-step-by-step-explanation-44e9eb85bf21>. Accessed: 2020-06-19.
- [15] Instagram. <https://it.wikipedia.org/wiki/Instagram>. Accessed: 2020-06-05.
- [16] Intuitive guide to latent dirichlet allocation. <https://towardsdatascience.com/light-on-machine-learning-intuitive-guide-to-latent-dirichlet-allocation-437c81220158>. Accessed: 2020-06-19.
- [17] Le 5v dei big data: dal volume al valore. https://blog.osservatori.net/it_it/le-5v-dei-big-data. Accessed: 2020-06-07.
- [18] Topic modeling using latent dirichlet allocation(lda) and gibbs sampling explained! <https://medium.com/analytics-vidhya/topic-modeling-using-lda-and-gibbs-sampling-explained-49d49b3d1045>. Accessed: 2020-06-29.
- [19] Tumblr. <https://it.wikipedia.org/wiki/Tumblr>. Accessed: 2020-06-05.
- [20] Tutto il valore dei big data: cosa sono e perché sono così importanti! https://blog.osservatori.net/it_it/big-data-cosa-sono. Accessed: 2020-06-06.
- [21] Twitter. <https://it.wikipedia.org/wiki/Twitter>. Accessed: 2020-06-05.

- [22] Understanding lstm networks. <https://colah.github.io/posts/2015-08-Understanding-LSTMs/>. Accessed: 2020-06-19.
- [23] Vine. [https://it.wikipedia.org/wiki/Vine_\(software\)](https://it.wikipedia.org/wiki/Vine_(software)). Accessed: 2020-06-05.
- [24] What is tf-idf? <https://monkeylearn.com/blog/what-is-tf-idf/>. Accessed: 2020-06-19.
- [25] Loris Belcastro, Riccardo Cantini, Fabrizio Marozzo, Domenico Talia, and Paolo Trunfio. Learning political polarization on social media using neural networks. *IEEE Access*, 8(1):47177–47187, 2020.
- [26] Gema Bello-Orgaz, Jason J. Jung, and David Camacho. Social big data: Recent achievements and new challenges. *Information Fusion*, 28:45 – 59, 2016.
- [27] Nada Ben-Lhachemi and El Habib Nfaoui. Using tweets embeddings for hashtag recommendation in twitter. *Procedia Computer Science*, 127:7–15, 01 2018.
- [28] Jianpeng Cheng, Li Dong, and Mirella Lapata. Long short-term memory-networks for machine reading. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 551–561, Austin, Texas, November 2016. Association for Computational Linguistics.
- [29] Ferraccioli Federico. *Topic Modeling, dietro le quinte: modelli grafici diretti e indiretti*. Tesi di laurea, Università degli studi di Padova, 2015/2016.
- [30] Frédéric Godin, Viktor Slavkovikj, Wesley De Neve, Benjamin Schrauwwen, and Rik Van de Walle. Using topic models for twitter hashtag recommendation. 05 2013.
- [31] Yang Li, Ting Liu, Jingwen Hu, and Jing Jiang. Topical co-attention networks for hashtag recommendation on microblogs. *Neurocomputing*, 331, 11 2018.
- [32] Eriko Otsuka, Scott Wallace, and David Chiu. A hashtag recommendation system for twitter data streams. *Computational Social Networks*, 3, 12 2016.

- [33] Tengyu Ma Sanjeev Arora, Yingyu Liang. A simple but tough-to-beat baseline for sentence embeddings. 2017.
- [34] Francesco Trombi. *Metodi per il Topic Detection su Twitter*. Tesi di laurea, Alma Mater Studiorum - Università di Bologna, 2016/2017.
- [35] Dongyao Wu, Sherif Sakr, and Liming Zhu. *Big Data Programming Models*. 02 2017.
- [36] Guixian Xu, Yueling Meng, Zhan Chen, Xiaoyu Qiu, Changzhi Wang, and Haishen Yao. Research on topic detection and tracking for online news texts. *IEEE Access*, PP:1–1, 04 2019.