Topic Modeling Assignment

A. Problem Statement

As part of the selection process we ask applicants to complete an assignment using their NLP and python skills. Assignment is to do the topic modeling of SEC 10Q/10K documents and extract the geographic segmentation data.

B. Details

Some information about 10K/10Q documents and geographic segmentation data:

a. 10K/10Q documents are quarterly filings of US public companies to SEC – securities regulatory body in US

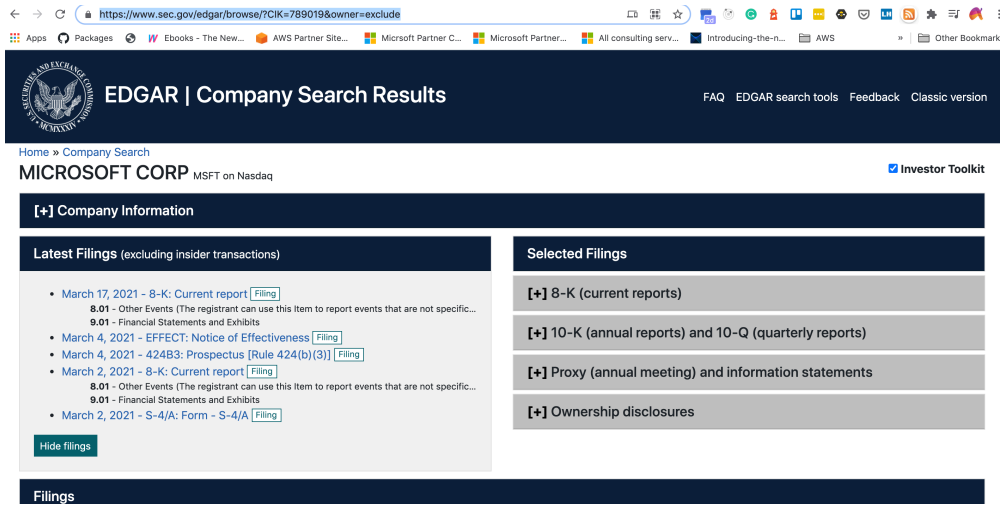
b. Geographic segmentation is information on companies' revenues from different regions of the world. For example,

You can Microsoft's revenue geographic segmentation can be found by following below steps –

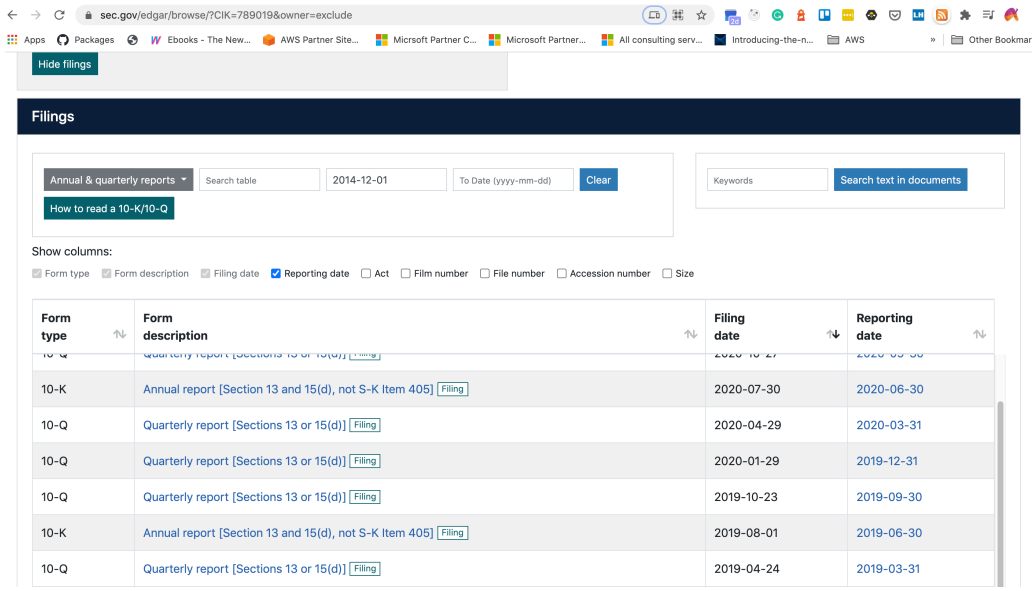
1. SEC link to search for filings of any company – https://www.sec.gov/edgar/searchedgar/companysearch.html

2. Microsoft 10K/10Q filings link –

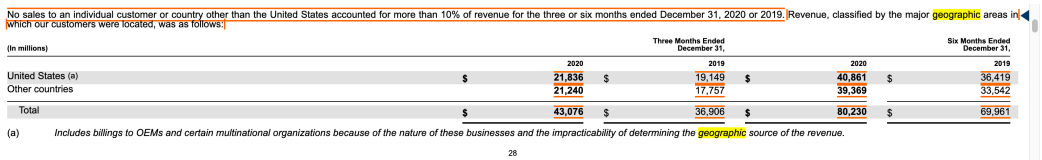
https://www.sec.gov/edgar/browse/?CIK=789019&owner=exclude



Select “View All 10K and 10Q” on the right



3. Select a 10Q report, and search for keyword – “segmentation” or “geographic segmentation” or “geographic”



No sales to an individual customer or country other than the United States accounted for more than 10% of revenue for the three or six months ended December 31, 2020 or 2019. Revenue, classified by the major geographic areas in which our customers were located, was as follows:

(In millions)	Three Months Ended December 31,		Six Months Ended December 31,	
	2020	2019	2020	2019
United States (a)	$ 21,836	$ 19,149	$ 40,861	$ 36,419
Other countries	21,240	17,757	39,369	33,542
Total	$ 43,076	$ 36,906	$ 80,230	$ 69,961

(a) *Includes billings to OEMs and certain multinational organizations because of the nature of these businesses and the impracticability of determining the geographic source of the revenue.*

28

C. Assignment

1. Develop model to extract geographic segmentation data from the 10K/10Q filings given a company name.
2. Use below link as a reference
    a. https://blog.quant-quest.com/using-topic-modelling-to-analyse-10-k-filings-notebook/
3. You are not required to use the topic modeling approach in #2, you can use other approaches if it yields better results
4. Required output is geographic segmentation data table as shown in B.b.3

D. Deliverables:

1. One-page document detailing the approach, steps and ML/NLP algorithms you would use to come up with the solution
2. Executable python code for the solution
3. CSV file with the features extracted using the solution
4. You can share github location to the solution

Deadline for the assignment: 25/03/2021

Please do submit the solution even if it is not complete end state solution, we are looking for your approach and application of skills to the problem solving

Reply back to this message with “I’m up for it” if you accept the assignment and we can expect a submission from you.

Please let us know if you have any questions or concerns

E. References:

1. https://blog.quant-quest.com/using-topic-modelling-to-analyse-10-k-filings-notebook/
2. https://rstudio-pubs-static.s3.amazonaws.com/163179_16606fecc68d46e4bec6d1c1a159b7c3.html
3. https://towardsdatascience.com/deep-learning-for-specific-information-extraction-from-unstructured-texts-12c5b9dceada
4. https://nanonets.com/blog/table-extraction-deep-learning/
5. Please google and download below papers for more background.

Automatic Tabular Data Extraction and Understanding

Roya Rastan

A thesis in fulfillment of the requirements for the degree of

Doctor of Philosophy

UNSW
AUSTRALIA

School of Computer Science and Engineering

Faculty of Engineering

The University of New South Wales

January 2017

Neural Topic Model with Reinforcement Learning

Lin Gui[1,§], **Jia Leng**[2,§], **Gabriele Pergola**[1], **Yu Zhou**[1], **Ruifeng Xu**[2,3,4], **Yulan He**[1,†]
[1]Department of Computer Science, University of Warwick, UK
[2]Harbin Institute of Technology (Shenzhen), China
[3]Peng Cheng Laboratory, Shenzhen, China
[4]Joint Lab of Harbin Institute of Technology and RICOH
Lin.gui@warwick.ac.uk, lengjia@stu.hit.edu.cn
gabriele.pergola@warwick.ac.uk, Yu.Zhou.1@warwick.ac.uk
xuruifeng@hit.edu.cn, yulan.he@warwick.ac.uk

Abstract

In recent years, advances in neural variational inference have achieved many successes in text processing. Examples include neural topic models which are typically built upon variational autoencoder (VAE) with an objective of minimising the error of reconstructing original documents based on the learned latent topic vectors. However, minimising reconstruction errors does not necessarily lead to high quality topics. In this paper, we borrow the idea of reinforcement learning and incorporate topic coherence measures as reward signals to guide the learning of a VAE-based topic model. Furthermore, our proposed model is able to automatically separating background words dynamically from topic words, thus eliminating the pre-processing step of filtering infrequent and/or top frequent words, typically required for learning traditional topic models. Experimental results on the 20 Newsgroups and the NIPS datasets show superior performance both on perplexity and topic coherence measure compared to state-of-the-art neural topic models.

1 Introduction

Probabilistic topic models have been used widely in nature language processing (Li et al., 2016; Zeng et al., 2018). The fundamental principle is that words are assumed to be generated from latent topics which can be inferred from data based on word co-occurrence patterns (Neal, 1993; Andrieu et al., 2003). In recent years, Variational Autoencoder (VAE) has been proved more effective and efficient to approximating deep, complex and underestimated variance in integrals (Kingma and Welling, 2013; He et al., 2017). However, the VAE-based topic models focus on the construction of deep neural networks to approximate the intractable distribution between observed words and latent topics based on log-likelihood and the learning objective is to minimise the error of reconstructing the original documents based on the learned latent topic vectors rather than improving the quality of learned topics, for example, measured by coherence scores (Kingma and Welling, 2013; Sønderby et al., 2016; Miao et al., 2016; Card et al., 2017; Srivastava and Sutton, 2017; Bouchacourt et al., 2018). The lack of consideration of topic coherence measures during the learning process of VAE-based topic models makes it difficult to control the quality of the generated topics. Intuitively, one solution is to jointly consider coherence scores in the learning objective. However, this is not feasible since coherence score is an unsupervised measure of topics based on a large-scale knowledge source, there is no ground truth "best topics".

Another limitation of existing approaches is that they typically require a pre-processing step to filter infrequent and/or top frequent words in order to reduce the vocabulary size and achieve better topic extraction results. Word filtering is often done heuristically. Although there have been attempts to automatically distinguishing background words and topic words, existing approaches either require a switch variable defined at each word position to indicate whether the word is a background word, which makes the models cumbersome, or model each latent topic as the deviation in log-frequency from a constant background distribution (Eisenstein et al., 2011; Smith et al., 2018).

In this paper, we propose a new framework to use reinforcement learning (Pan et al., 2018; Qin et al., 2018; Yin et al., 2018) to incorporate the topic coherence measures into the learning of a neural topic model and filter background words dynamically. More concretely, given an input document, its constituent words will first be sampled

[§]The two authors contributed equally to this work.
[†]Corresponding author.

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 3478–3483, Hong Kong, China, November 3–7, 2019.

Neural Topic Modeling with Continual Lifelong Learning

Pankaj Gupta [1] **Yatin Chaudhary** [1 2] **Thomas Runkler** [1] **Hinrich Schütze** [2]

Abstract

Lifelong learning has recently attracted attention in building machine learning systems that continually accumulate and transfer knowledge to help future learning. Unsupervised topic modeling has been popularly used to discover topics from document collections. However, the application of topic modeling is challenging due to data sparsity, e.g., in a small collection of (short) documents and thus, generate incoherent topics and sub-optimal document representations. To address the problem, we propose a lifelong learning framework for neural topic modeling that can continuously process streams of document collections, accumulate topics and guide future topic modeling tasks by knowledge transfer from several sources to better deal with the sparse data. In the lifelong process, we particularly investigate jointly: (1) sharing generative homologies (latent topics) over lifetime to transfer prior knowledge, and (2) minimizing catastrophic forgetting to retain the past learning via novel selective data augmentation, co-training and topic regularization approaches. Given a stream of document collections, we apply the proposed Lifelong Neural Topic Modeling (LNTM) framework in modeling three sparse document collections as future tasks and demonstrate improved performance quantified by perplexity, topic coherence and information retrieval task. Code: https://github.com/pgcool/Lifelong-Neural-Topic-Modeling

1. Introduction

Unsupervised topic models, such as LDA (Blei et al., 2003), RSM (Salakhutdinov & Hinton, 2009), DocNADE (Lauly et al., 2017), NVDM (Srivastava & Sutton, 2017), etc. have been popularly used to discover topics from large document collections. However in sparse data settings, the application of topic modeling is challenging due to limited context in a small document collection or short documents (e.g., tweets, headlines, etc.) and the topic models produce incoherent topics. To deal with this problem, there have been several attempts (Petterson et al., 2010; Das et al., 2015; Nguyen et al., 2015; Gupta et al., 2019) that introduce prior knowledge such as pre-trained word embeddings (Pennington et al., 2014) to guide meaningful learning.

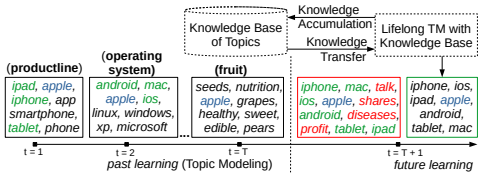


Figure 1. Motivation for Lifelong Topic Modeling

Lifelong Machine Learning (LML) (Thrun & Mitchell, 1995; Mitchell et al., 2015; Hassabis et al., 2017; Parisi et al., 2019) has recently attracted attention in building adaptive computational systems that can continually acquire, retain and transfer knowledge over life time when exposed to modeling continuous streams of information. In contrast, the traditional machine learning is based on isolated learning i.e., a one-shot task learning (OTL) using a single dataset and thus, lacks ability to continually learn from incrementally available heterogeneous data. The application of LML framework has shown potential for supervised natural language processing (NLP) tasks (Chen & Liu, 2016) such as in sentiment analysis (Chen et al., 2015), relation extraction (Wang et al., 2019), text classification (de Masson d'Autume et al., 2019), etc. Existing works in topic modeling are either based on the OTL approach or transfer learning (Chen & Liu, 2014) using stationary batches of training data and prior knowledge without accounting for streams of document collections. The unsupervised document (neural) topic modeling still remains unexplored regarding lifelong learning.

In this work, we explore unsupervised document (neural) topic modeling within a continual lifelong learning paradigm to enable knowledge-augmented topic learning over lifetime. We show that *Lifelong Neural Topic Modeling*

[*]Equal contribution [1]Corporate Technology, Siemens AG Munich, Germany [2]CIS, University of Munich (LMU) Munich, Germany. Correspondence to: Pankaj Gupta <pankaj.gupta@siemens.com>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 108, 2020.