Analisis de Datos en Amazon Web Services (AWS)

Andrea Villanes, Ph.D.
Assistant Professor – Institute for Advanced Analytics

Email: avillan@ncsu.edu
LinkedIn: https://www.linkedin.com/in/andreavillanes/
GitHub: https://github.com/andreavillanes
Twitter: @andrea.villanes

Background Information

Education:

- 2019 Ph.D. in Computer Science, NCSU
- 2012 M.Sc. in Computer Science, NCSU
- 2009 M.Sc. in Analytics, NCSU

Work Experience:

- 2019 –Assistant Professor
- 2013 – Research Associate
- 2012/2013 – Inmar Marketing Insights Intern/Extern
- 2010/2011 – Walmart Marketing insights Intern/Extern

www.menti.com

Agenda

1. Text Mining: Analisis de Textos
2. Sentiment Analysis: Sentimiento de Datos
3. Amazon Web Services (AWS)
4. Demo de AWS

Text Mining – What is it?

"...find interesting regularities in large textual datasets" (Fayad)

donde interesante significa no-trivial, escondido, desconocido, y potencialmente util.

Text Mining – Process Objective

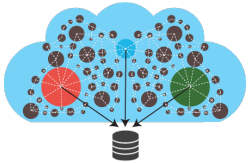
Flight 2463 leaving West Palm Beach (PBI) at 2.42pm on June 15 arriving at New York LaGuardia (LGA) at 5.30pm. I was slated to take
Delta Website froze 4 times trying to set up flight to three different locations. Had to call and set up over the phone only to find out
I just returned from a round-trip First/Business Elite FLL-ATL-MUC-ATL-FLL trip. I really must say that all the flights were great. Every
Round-trip flight from Quito Ecuador to Birmingham Alabama with one stop in Atlanta. The check in process was difficult since we ar
Narita - Bangkok June 13 Business Elite 747. Slight delay due to hold for other delayed inbound flights but all in all an excellent flight.
Flight from NY La Guardia to Cleveland OH at 01.15am email said flight would be delayed by 2 hours. Delayed our taxi to airport by 2
Originally I had a 2 hour layover. Delta changed our flights to a 4 hour layover. The plane now has been delayed another 20 minutes
We flew paid business class fares in Delta's Business Elite LAX to SYD and SYD to LAX in the forward B/E cabin on Delta's 777-200LR.
I flew from Heathrow to Seattle. To be honest I was dreading the flight as I had read some really bad reviews of Delta. But it was exc
I was a bit stubborn about flying Delta for the first time since I'm a fan of Frontier and I don't fly very often about once every three t
JFK-LHR. Had a great on time flight out of JFK with Delta. I found the service friendly and food was great. Constantly coming through
My wife and I fly frequently and over the last couple of years have found ourselves on Delta a couple of times. Each of the prior trips
DL 1134 PBI-ATL. Great ground experience. Inbound flight late prior to equipment arrival gate agents asked for volunteers to check b
Our flight from Fairbanks to Minneapolis was a great flight with great service. On our return flights the service was great. However th
On May 22 after a 6-hour delay 5 alerts and us waiting with hope at the terminal flight 5030 to Atlanta was canceled. And as per airp
Considering how Delta Airlines was an American Air carrier I initially wasn't expecting much and imagined that service would be medi
Travelled MSP-LHR AMS-MSP in May 2014. Flights were on time. Flight attendants responsive enough and food was okay. Economy c
JFK-LAX on a 757-200 in Business on May 17 angled seats with ample leg room excellent service then onward to HNL on a 757-300 in
Third long haul flight with Delta LHR-ATL-TPA-ATL-LHR all flights on time first transatlantic flight they run out of food selections so or

	Term 1	Term 2	Term 3	Term 4	Term 5	…	Term n-1	Term n
Doc 1	3	0	5	3	0	2	3	0
Doc 2	4	5	6	0	3	4	0	1
Doc 3	0	2	2	3	0	4	2	0
Doc 4	1	2	0	2	5	0	4	3
…								
Doc n-2	7	2	2	3	0	0	1	1
Doc n-1	7	2	0	0	0	3	1	0
Doc n	0	2	3	3	4	0	1	1

The overlap between term vectors might provide some clues about the similarity between documents.

Text Mining – Transformation Process

Text Mining – Data Collection



Web crawling: Internet bot that systematically browses the World Wide Web, typically for the purpose of Web indexing (web spidering).

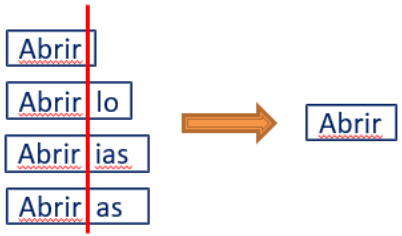
APIs: Twitter, Trip Advisor, Yelp, LinkedIn

CSV files: surveys, emails, open ended questions… and more.

Text Mining – Text Parsing

Text Cleaning: remove unnecesary words

Stop Words: remover palabras comunes pero que no proveen utilidad al descubrimiento del contexto (el, la, de, los, y, etc...)



Stemming: convierte las palabras a su raíz.

Text Mining – Term Weighting

Term Frequency–Inverse Document Frequency (TF-IDF)

$$\text{TF-IDF}_{i,j} = f_{i,j} \times \log\left(\frac{n}{n_i}\right)$$

TF-IDF gives greater weight to terms that occur more frequently within a document (TF), but infrequently across the document collection (IDF). Intuitively, TF-IDF implies that if a term t_i occurs frequently in a document D_j, it is an important term for characterizing D_j .Moreover, if t_i does not occur in many other documents, it is an important term for distinguishing D_j from other documents

Text Mining – Process Objective

	Term 1	Term 2	Term 3	Term 4	Term 5	…	Term n-1	Term n
Doc 1	3	0	5	3	0	2	3	0
Doc 2	4	5	6	0	3	4	0	1
Doc 3	0	2	2	3	0	4	2	0
Doc 4	1	2	0	2	5	0	4	3
…								
Doc n-2	7	2	2	3	0	0	1	1
Doc n-1	7	2	0	0	0	3	1	0
Doc n	0	2	3	3	4	0	1	1

- Clustering (segmentacion)
- Clasificacion
- Associacion de palabras

Text Mining – Tools

	SAS Enterprise Miner (Text Miner) SAS Viya → Text parsing, Term weighting, LSA, modeling
	Needed <- c("tm", "SnowballCC", "RColorBrewer", "ggplot2", "wordcloud", "biclust", "cluster", "igraph", "fpc") → Text parsing, term weighting, LSA, LDA, NMF, modeling
	scikit-learn, nltk, numpy, pandas, beautiful soup → Web crawling, text parsing, term weighting, LSA, LDA, NMF, modeling

Agenda

1. Text Mining: Analisis de Textos
2. Sentiment Analysis: Sentimiento de Datos
3. Amazon Web Services (AWS)
4. Demo de AWS

Sentiment Analysis

Sentiment analysis is the computational task of determining the attitude towards a target or topic in a piece of text.

La actitud se puede definir como un juicio evaluativo (positivo-negativo) o como un juicio emocional (por ejemplo: frustración, alegría, enojo, tristeza)

Sentiment Analysis – Sentiment Dictionaries

Sentiment dictionaries provide prior associations of words and phrases with their respective annotated sentiment.

Examples:

- Affective Norms for English Words (ANEW) was built to assess the emotional affect for a set of verbal terms. Three emotional dimensions were scored: valence (or pleasure), arousal (or activation), and dominance. A total of 1,033 word that were previously identified as emotion-carrying words, and that provided good coverage of all three dimensions, were rated.
- SentiStrength was developed from manually scoring social media comments on two five-point scales representing both the positive and the negative emotion of a comment.
- EmoLex was developed using Amazon Mechanical Turk. EmoLex focused on the eight emotions defined by Plutchik: joy, sadness, anger, fear, trust, disgust, surprise, and anticipation. In addition, the lexicon provides valence scores. EmoLex has 14,182 unigrams (words), and includes most frequent English nouns, verbs, adjectives, and adverbs.

Agenda

1. Text Mining: Analisis de Textos
2. Sentiment Analysis: Sentimiento de Datos
3. Amazon Web Services (AWS)
4. Demo de AWS

Cloud Computing – What is it?

Cloud computing is the on-demand delivery of compute power, database, storage, applications, and other IT resources via the internet with pay-as-you-go pricing.

Cloud computing gives you access to servers, storage, databases, and a broad set of application services over the Internet. A cloud services provider such as Amazon Web Services, owns and maintains the network-connected hardware required for these application services, while you provision and use what you need via a web application.

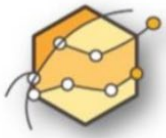
Ventajas de Cloud Computing

Trade capital expense for variable expense.

Increase speed and agility.

Benefit from massive economies of scale.

Stop spending money on running and maintaining data centers.



Stop guessing capacity.

Go global in minutes.

What is AWS?

AWS is a collection of remote computing services called web services. These web services make up a cloud computing platform offered via the internet. AWS delivers web-based cloud services for storage, computing, networking, databases, and more.

Top players:

aws

Microsoft Azure

Google Cloud Platform

Agenda

1. Text Mining: Analisis de Textos
2. Sentiment Analysis: Sentimiento de Datos
3. Amazon Web Services (AWS)
4. Demo de AWS

Utilizando Python para text mining....

```
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.model_selection import cross_val_score

# build BOW features on train articles
cv = CountVectorizer(binary=False, min_df=0.0, max_df=1.0)
cv_train_features = cv.fit_transform(train_corpus)
```

```
# transform test articles into features
cv_test_features = cv.transform(test_corpus)
```

```
print('BOW model:> Train features shape:', cv_train_features.shape, ' Test features shape:', cv_test_features.shape)
```

```
BOW model:> Train features shape: (12263, 66865)  Test features shape: (6041, 66865)
```

```
from sklearn.naive_bayes import MultinomialNB

mnb = MultinomialNB(alpha=1)
mnb.fit(cv_train_features, train_label_names)
mnb_bow_cv_scores = cross_val_score(mnb, cv_train_features, train_label_names, cv=5)
mnb_bow_cv_mean_score = np.mean(mnb_bow_cv_scores)
print('CV Accuracy (5-fold):', mnb_bow_cv_scores)
print('Mean CV Accuracy:', mnb_bow_cv_mean_score)
mnb_bow_test_score = mnb.score(cv_test_features, test_label_names)
print('Test Accuracy:', mnb_bow_test_score)
```

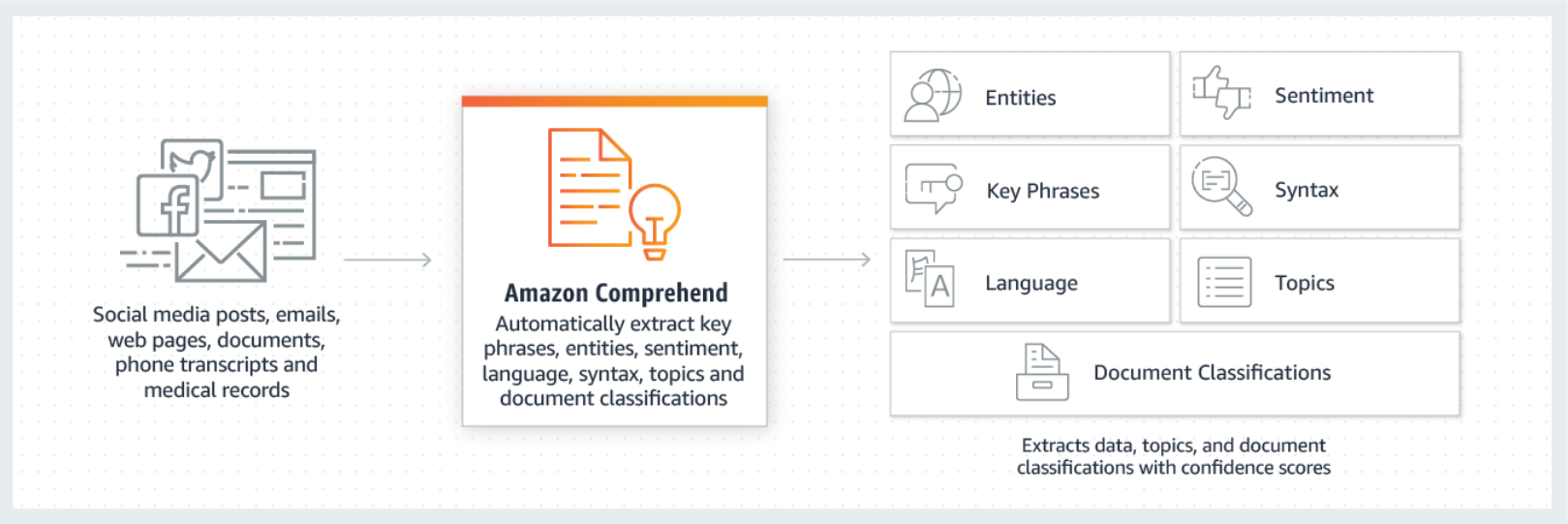
```
CV Accuracy (5-fold): [ 0.68468102  0.68241042  0.67835304  0.67741935  0.6792144 ]
Mean CV Accuracy: 0.680415648396
Test Accuracy: 0.680185399768
```

```
from sklearn.linear_model import LogisticRegression

lr = LogisticRegression(penalty='l2', max_iter=100, C=1, random_state=42)
lr.fit(cv_train_features, train_label_names)
lr_bow_cv_scores = cross_val_score(lr, cv_train_features, train_label_names, cv=5)
lr_bow_cv_mean_score = np.mean(lr_bow_cv_scores)
print('CV Accuracy (5-fold):', lr_bow_cv_scores)
print('Mean CV Accuracy:', lr_bow_cv_mean_score)
lr_bow_test_score = lr.score(cv_test_features, test_label_names)
print('Test Accuracy:', lr_bow_test_score)
```

Amazon Comprehend

Amazon Comprehend



Pasos a seguir

1. La data (Quejas_Aerolineas.csv) se encuentra en Moodle
2. Subir la data que queremos analizar a AWS S3
3. Analisis de textos, AWS Comprehend, dos opciones:
    1. Analisis de documento por document
    2. Analisis de varios documentos al mismo tiempo

Muy decepcionado con Delta Airlines. Intente planificar y hacer un presupuesto para todo y me sentí muy molesto cuando registré dos maletas por $ 25 cada una. Delta anuncia en la parte posterior de su boleto "Su primera maleta documentada es siempre gratis" (dentro del tamaño recomendado y con menos de 50 libras revisé su sitio web a nuestro viaje para evitar cargos). No me importa si me cobran por mi equipaje, pero en serio, no anuncien una cosa y digan otra. Si reserva sus boletos con 6 meses de anticipación y elige sus asientos en ese avión y el avión está vacío cuando los elige, al menos debería poder sentarse junto a su pareja, pero no se moleste, es una pérdida de tiempo. . Satisfecho con la tripulación de vuelo y los asistentes de vuelo, felicitaciones por sus actitudes positivas. Sin embargo, no creo que vaya a volar con Delta pronto.

Gracias!

Email: avillan@ncsu.edu
LinkedIn: https://www.linkedin.com/in/andreavillanes/
GitHub: https://github.com/andreavillanes
Twitter: @andrea.villanes