




Text Analytics Fundamentals

Cleaning and treating text, creating document vectors, and modeling using TF-IDF

Agenda

- Fundamentals
 - Tokens and terms
 - Dictionaries and document vectors
 - Stemming and lemmatization
- Term Frequency (TF) and Inverse Document Frequency (IDF)
 - Creating an inverted index and retrieving documents from a query

Structured vs Unstructured Data

- Structured – tabular data
 - Semi-structured – non-tabular data with some meta-data
 - Example: JSON, XML
 - Unstructured – non-tabular data with no predefined schema
- 

Structured – Tabular Data

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin
1	1	0	3	Braund, Mr. Owen Harris	male	22.00	1	0	A/5 21171	7.2500	
2	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38.00	1	0	PC 17599	71.2833	C85
3	3	1	3	Heikkinen, Miss. Laina	female	26.00	0	0	STON/O2. 3101282	7.9250	
4	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.00	1	0	113803	53.1000	C123
5	5	0	3	Allen, Mr. William Henry	male	35.00	0	0	373450	8.0500	
6	6	0	3	Moran, Mr. James	male	NA	0	0	330877	8.4583	
7	7	0	1	McCarthy, Mr. Timothy J	male	54.00	0	0	17463	51.8625	E46
8	8	0	3	Palsson, Master. Gosta Leonard	male	2.00	3	1	349909	21.0750	
9	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.00	0	2	347742	11.1333	
10	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.00	1	0	237736	30.0708	
11	11	1	3	Sandstrom, Miss. Marguerite Rut	female	4.00	1	1	PP 9549	16.7000	G6
12	12	1	1	Bonnell, Miss. Elizabeth	female	58.00	0	0	113783	26.5500	C103
13	13	0	3	Saunderscock, Mr. William Henry	male	20.00	0	0	A/5. 2151	8.0500	
14	14	0	3	Andersson, Mr. Anders Johan	male	39.00	1	5	347082	31.2750	
15	15	0	3	Vestrom, Miss. Hulda Amanda Adolfina	female	14.00	0	0	350406	7.8542	

Semi-structured Data

```

1  <html>
2  <head>
3  <title>CSS Experiments</title>
4  <link rel="stylesheet" href="styles.css" type="text/css" media="all">
5  </head>
6  <body>
7  <div id="menu">
8  <ul>
9      <li><a href="http://abduzeedo.com/">Home</a></li>
10     <li><a href="http://abduzeedo.com/tutorials">Tutorials</a></li>
11     <li><a href="http://abduzeedo.com/tags/interview">Interviews</a></li>
12     <li><a href="http://abduzeedo.com/tags/wallpaper">Wallpapers</a></li>
13 </ul>
14 <input type="" name="" value="" />
15 </div>
16 <div id="flickr_badge_uber_wrapper">
17     <div id="flickr_badge_wrapper">
18         <script type="text/javascript" src="http://www.flickr.com/
            badge_code_v2.gne?
            count=12&display=latest&size=s&layout=x&source=user_set&user=764
            66518%40N00&set=72157604672645588&context=in
            %2Fset-72157604672645588%2F"></script>
19     </div>
20 </div>
21
22 </body>
23 </html>|

```

Unstructured Data



TIME  @TIME · 52s



An earlier version of this story incorrectly stated that the National Weather Service mistakenly sent a tsunami warning to phones. The warning was sent by third-party weather apps, not by the National Weather Service. The tweet was since deleted



A Tsunami Warning Blared on Phones Across the Country This Morni...

"Please note there is NO TSUNAMI THREAT"

time.com