Scope Proposal - USC/ThreadTogether Spring 2020 Text Analytics

Contact: Karin Chu, Chief Data & Analytics Officer
- LinkedIn
- Email

I. Project Proposal

There are 2 main sources of input data that Thread Together (TT) leverages – Image and web data. Our present proposal pertains to mining the web data via text analytics.

There are **4 primary objectives:**

A. **Attribution of web data** – TT will provide a dataset previously extracted from various retail websites, we would like to organize and extract meaningful product information from the web data.

B. **Attribution tool/app** – As a supplement to extracting features from current datasets, we propose the development of a text parsing app that would provide a repeatable solution and will require minimal human intervention for future datasets from retailers. The input to this app would be product descriptions, tags, and other metadata.

C. **Business application** – It is our hope that by immersing the students in solving a real world business problem with active business/technical support, that the students will gain immediate and valuable experience as their competitive edge in the evolving data science market.

D. **Outfitting recommendation** – Given tagged data about different women's fashion items, produce a customer recommendation system.

II. Data

This section describes the available data and the supporting library of attributes.

A. Leveraging previously scraped web data, TT product data consist of mainly 2 forms of text (verbatim to reflect actual nature of data):

1. Product descriptions – "*People often think of sundresses when temperatures rise, but Zimmermann's floral-print 'Amari' jumpsuit is an equally good alternative. Its ruched sleeveless bodice is designed with tie shoulder straps and a square neckline with shirred trim. Wear the wide-leg cotton style with flat sandals and oversized gold jewelry*. "

2. Product details –
Zip fastening along back
Composition: 100% cotton
Dry clean
Imported

This style fits true to size-We suggest taking your normal size

B. The objective of our initiative is **focused on women's fashion.** However, the data sources may also contain non-relevant products depending on the brand and retail sites.

Examples of non-relevant products: Home goods, holiday decorations, etc.

C. There are 2 main types of attributes to identify:

1. **Explicit product attributes** – Color, material, care, sleeve length, etc.
2. **Implicit product attributes** – Business casual, season, etc.

D. TT has created a library of pre-determined fashion attribute names and a list of possible values; the students may leverage this information to shave off time required.

E. For any additional attribute names/values developed through this project, we also ask that the new attribute names/values be included in the library. The categorization of the terms in the library will be optional and can be supported internally by the TT Data Science team if need be.

III. Methods

We recommend the following flow from a business end-user perspective, please adjust as necessary:

A. **Identifying relevant fashion products from non-relevant products.** We have put this as the first step based on natural data constraints from the brands (II.B).

B. Of the relevant fashion products, using the product description & product detail with the TT library of terms to identify (II.A, C, D):

1. **Explicit product attributes**
2. **Implicit product attributes.**

C. **Identifying any gaps in attributes and document** in the attribution library (II.E). Gaps can come in 2 forms:

1. Missing because the attributes will require images, or, less common,
2. The attributes are entirely missing which indicates that we need broader product selections for coverage. Example: Lack of ¾ sleeves

D. **Packaging code into a repeatable solution** such as an app that would identify attribute values based on product descriptions with minimal user intervention, or a priori tagging

E. **Outfit recommendation** based on tagged attributes.