

Text Analytics for Collaborative Content (TACC)

Introduction

Collaboration is the key aspect in every organization where opinions of their policies and programs (or products and services) are collected from wide variety of custodians like : the society (or customers), policy stakeholders (or employees) and experts (business partners).

Organization face challenges in terms of analyzing and reporting content coming from various mediums including but not limited to Feedbacks, Reviews, Comments, Social Media Interactions, Blogs and Surveys that are hosted internally or on the cloud. Manual evaluation of these are quite impossible and there is a need to automate the review process and bring out the message in a format which stakeholders can understand and act upon appropriately. We are living in information age and the content is getting generated through various sources every second. Sourcing all the unstructured information into a common platform and processing are the challenges faced by the Data Engineers. Named Entity Recognition (NER), Relationship Extraction , Sentiment Analysis (Polarity) are key aspects of a Text Analysis Process. Many research institutions are contributing heavily into this platform to enable the machine learning from the human generated content.

Business Scenario

ASSYST has built a platform for online collaboration channels that can enable the features an organization would like to activate for Analysis. The current version of the collaboration engine supports six channels for data collection; Review / Feedbacks , Comments , Social Media Interactions (LinkedIn, Facebook, Twitter and Google+), Blogs and Surveys. These channels act as data consumers and store those in data repositories for analysis. This also allows submissions of additional materials pertains to any channels; which allows content extraction from documents (PDF, Excel and Documents) as well.

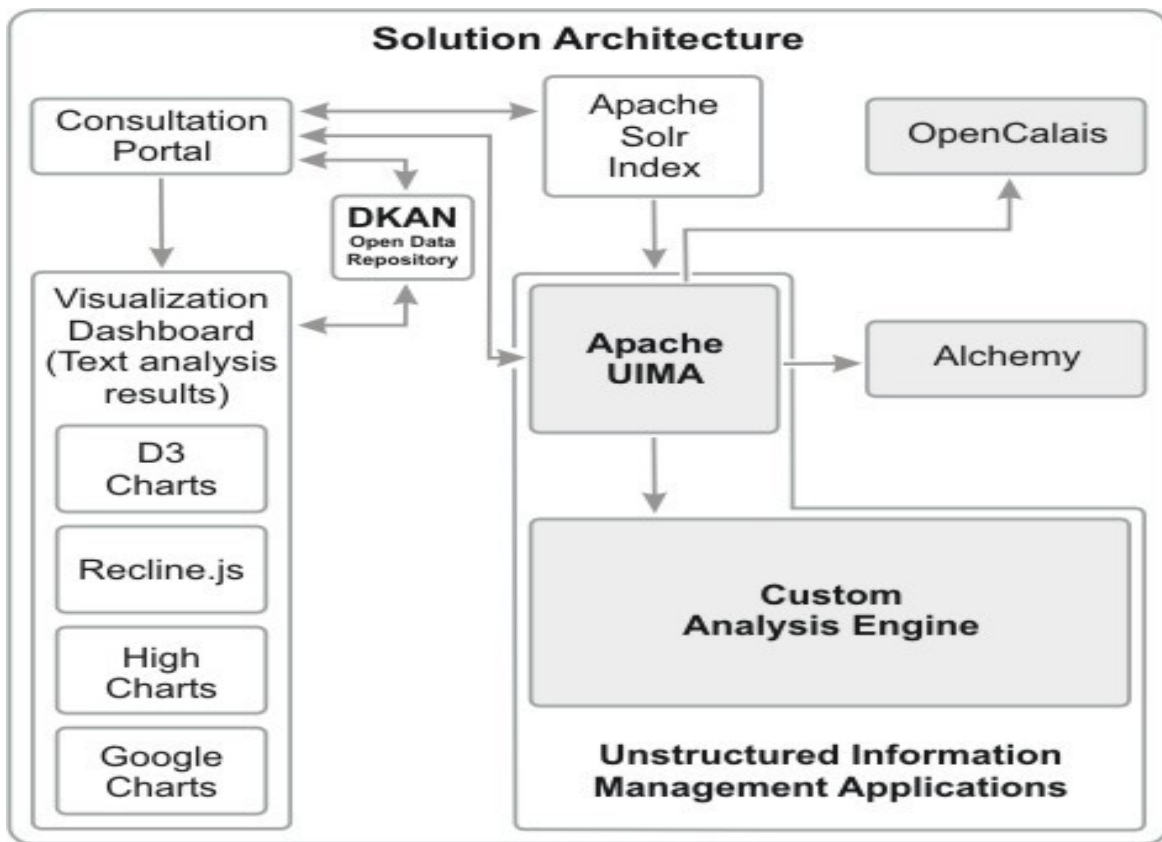
Collaboration Hub is the platform where all the individual consultation specific analysis gets added. The solution enables the stakeholders to pick and choose the source of information (channels) depending on the interest or the need of the data collection practice. For example, someone wants to compare meeting comments versus social media comments. Otherwise, the user wants to filter analysis based on the geographical or sector or topic or theme based. The reason is that, when we compare the sentiments of two data sets if both are from different Sector or Topic or Geography; those results may be logically different . Technically the data may be correct, but contextually that may be incorrect. The



solution should enable the custodians to identify and define the logical relationships depends on the context of the business function .

Organizations using this tool can build their own vocabulary (Keywords , Entities and Concepts) for processing the data collected through various sources. The institutions follow specific vocabulary to avoid confusions or different interpretations on the results. This tool also facilitates integration with industry proven data analysis platforms like Alchemy and Open Calais along with Assyst Custom Analysis Engine. The custom solution benefits advanced optimization to narrow down the results and fine tune the process according to the business requirements.

Solution Architecture



Picture 2 : Solution Architecture

Our experience in collecting, processing and managing the open data; accelerated our thoughts and actions to arrive an apt solution to this business requirement. The core ideology was to reuse the solution stack as much possible and remain on the open space to avoid additional financial burden. We were successful in that to a great extend; Solr based Apache UIMA (Unstructured Information Management Applications) solution has been identified. It is an API which connects easily with Solr and allows connectivity with various open text analysis engines (Alchemy and OpenCalais). It was a successful story and we were able to gather Keywords, Concepts and Entities out of the text passed to the analysis engine.

The critical aspect of the whole solution is that, how best we portray the processed information in a way it connects to all sort of people easily and effectively; through strong self depicting visualizations. Assyst has invented a strong portable data structure to enable any data visualization tools to consume

