Assignment 2: Working with XML Tags

Predict 453

Section 55

Spring Quarter

School of Continuing Studies

Northwestern University

Daniel Prusinski

In compliance for Master of Science Predictive Analytics

Bachelor of Science Business Marketing

Data Analyst

US Bank

220 S 6th St

Minneapolis, MN

In order to fulfill the requirements for this assignment, one must understand the basics of XML Tags. While this topic of XML Tags is new, by the end of this exercise a basic understanding will be demonstrated.

XML was the product of an eleven person team in 1998, and was developed to assuage issues as the prevalence of the internet grew (Kimber, 2006). The acronym stands for Extensible Markup Language and was developed by the World Wide Web Consortium (W3C) to create customized Tags that allowed: communication, metadata compilation, validation, and synchronization between users of data throughout the internet (Balas, 2002). Tagging directly pertains to this assignment, but one must first understand that the term Element and Tag are used interchangeably (Tizag.com). Tagging allows one to subdivide text into categories that make it easier to understand what type of text is being displayed. When a large amount of information is communicated, Tagging is a great way to organize and sort through the data (3schools.com). The important take-away from this brief introduction to XML is that Tagging allows one to subdivide data in a text environment.

Two text XML documents have been supplied to analyze, and at first-glance the XML Tagging provides a great meta-data snapshot. For both the "Q12012 Alternative BioTech Inc 91812" (Q1) document and "Q32011 Innovative Technology Inc 91812", (Q3) there is one major Tag 'document', and within this Tag all the other Tags are subset and so too the actual text. There is another Tag 'management' that contains just the initial presentation for the entire interview

text for (Q3), and this Tag is a 'child' Tag of the 'document' Tag. The grandchild Tag of 'document' and child Tag of 'management' contains the 'question' and 'answer' Tags for the interview. To recap, this document has three layers: the first, 'document' contains all the text, the second, 'management' contains the initial presentation, the third 'answer' and 'question' contain the remaining text for the document. Q1 on the other hand has the 'parent' Tag document and at the child level there are three Tags 'management', 'question', and 'answer'. This document only has two levels, which will mildly differentiate the querying from Q3.

Now that the tiers of the Tags have been defined, focusing on the actual Tags for content is the logical next step. From a macro view, searching similar Tags throughout multiple documents is a great was to put parameters on a search. If one were to myopically search the Tag 'management' between the documents, the results would focus on initial quantitative information. From searching this Tag one would gather quarterly figures as well as leadership's vision moving forward. If the whole document was analyzed, the results might be misleading in that depending on the text search misleading results might be generated. For example, if one searched for the phrase 'investment%(substring 1,50)' in just 'management' Tags one would have results like "investment in our manufacturing facilities in Q3 and we also". On the other hand, if the 'management' Tag was not used to delineate the search, results like "Investments. Do you expect to show a profit in the current q" would populate

the query and make the results more ambiguous. By focusing in on specific Tags it is easier to delineate context and save time wading through different results. On the other hand, searching the Tag 'question' will yield results like, "Erik Abrams, Capital Investment Strategies" and "Peter Hopkins, ABC Investments". This search is relevant for searching the titles for different people asking questions at the quarterly meeting. By searching specific Tags the context of a word can change drastically, which has been demonstrated by the simple example of searching the word 'investment' through different Tags. It is a rather obvious conclusion that searching the whole document will yield ambiguous results that are not veracious to the original search.

Logic can also be applied to searching different Tags. One would not expect to find critical or poignant remarks in the 'management' Tags, thus searching just the 'question' Tags for specific poignant words would yield more meaningful results. Searching the word 'profit' in the 'question' Tag will yield either a positive or negative feeling about the profits as well as the nature of the profits. Combining this search with the 'management' Tag will again yield ambiguous results based on the context of the word.

While the topic of XML Tags are new to me, this exercise has provided a meaningful demonstration and built my knowledge on the usefulness and necessity of utilizing Tags in preliminary text mining. Both the structure and string value of Tags provide advantageous opportunities to search, refine, and mine text documents.

Works Cited

Balas, Janet. "What Is This XML Thing and Why Do I Need to Know About It?." *Computer in Libraries* 22.8 (2002): 39-41. *EBSCO Host*. Web. 11 Apr. 2013.

Kimber, Eliot. "Dr. Macro's XML Rants: XML: Ten Year Aniversary." *Dr. Macro's XML Rants*. N.p., n.d. Web. 12 Apr. 2013. <http://drmacros-xml-rants.blogspot.com/2006/11/xml-ten-year-aniversary.html>.

"XML Elements." *W3Schools Online Web Tutorials*. N.p., n.d. Web. 12 Apr. 2013. <http://www.w3schools.com/xml/xml_elements.asp>.

"XML Tutorial - Element." *Tizag Tutorials*. N.p., n.d. Web. 12 Apr. 2013. <http://www.tizag.com/xmlTutorial/xmlelement.php>.

.