P: ISSN No. 0976-8602 RNI No.UPENG/2012/42622 VOL.-7, ISSUE-4, October-2018

E: ISSN No. 2349-9443 Asian Resonance

Resume Parsing: Digging Pearls from Sea

Shameen Warsi
Research Scholar,
Institute of Management Studies,
Devi Ahilya Vishwavidyalaya,
Indore

N.K.Totala
Reader,
Institute of Management Studies,
Devi Ahilya Vishwavidyalaya,
Indore

Abstract

Resume parser compiles and interprets the unstructured but important data found in resumes into structured form. Regular Expression / Rule Based Parsing is being utilized by many HR professionals. Latest resume parsing technology is based on Natural Language Processing (NLP). It utilizes the advanced techniques of named entity recognition and co-reference resolution to read the free text and provides unbelievable results near to human accuracy. This research paper explores, analysis and comparison of NLP based resume parsing tools helping in predictive analytics for HRM with the help of action research. It also flashes light on advancement of NLP based resume parsing over Regular Expression/Rule Based Parsing.

Keywords: Resume Parsing, Natural Language Processing (NLP), HRM, Predictive Analytics, Regular Expression/Rule Based Parsing, etc.

Introduction

The word meaning of parsing is to analyze, to uncover a deeper meaning, or to discover implications. Practically it also means to find desired things objectively which is hidden or covered.

Resume Parsing

Resumes are commonly presented in MS Word format; making it is easy for human beings to read and understand but the computers find it difficult to interpret and infer. The conversion of a regular CV/Resume into XML format or structured information to make it easy for analysis, understanding and reporting is termed as resume or CV parsing to eliminate the manual process of reviewing every resume/job application. It saves time, energy and money and speed up the task. A resume parser is a computer program that can analyze any document and extract the desired key information like, skills, education, qualifications, work experience, contact information, etc. (Daxtra, 2016). Resume parsing is a practice that converts unstructured form of resume data into the desired structured format. Resume parser as a software program analyses a resume/CV data and extracts into machine-readable output such as XML, JSON, etc. With the help of this software, it becomes easy to store and analyze resume data (Rchilli, 2016).

Types of Parsing

Regular Expression

A regular expression also abbreviated as ‘regex’ is a way for a computer user or programmer to express how a computer program should look for a specified pattern in text and then what the program is to do when each pattern match is found is to report accordingly. A regular expression could tell a program to search for all text lines that contain a specific word and then to print out each line in which a match is found or substitute another text sequence (TechTarget, 2006). A “regex” is a string of text that allows creating patterns that help in matching, locating, and managing text. *Perl* being an illustration of a programming language utilizes regular expressions. It is only one of the many places to find regular expressions. Regular expressions can also be used from the command line and in text editors to find text within a file. When first trying to understand regular expressions it seems as if it is a different language. Mastering regular expressions can save thousands of human hours if the work is with text or need to parse large amounts of data (Computer Hope, 2017).

Rule Based Parsing

Rule Based parser is a parser that uses hand written i.e., designed rules as opposed to rules that are derived from the data. Rule based parsing uses semantic rather than syntactic parsing. It can handle the documents in pdf, txt, doc, and docx format.

Natural Language Processing (NLP) Based Parsing

For more than 50 years study of natural language processing is being tried and done. NLP is a branch of artificial intelligence that helps computers in understanding, interpreting and manipulating human language/s. NLP draws its property from many disciplines such as computer science and computational linguistics, in its pursuit to bridge the gap between human communication and computer understanding and interpretation. NLP helps computers in communicating with humans in computers' language and scales out other language-related tasks and issues. NLP makes it possible to read text, hear speech, interpret them, measure sentiment and determine which parts are significantly important as per requirements. Contemporary machines can analyze more language-based data than humans, without fatigue and in a consistent and unbiased manner. Considering the staggering and huge amount of unstructured data generated on day to day basis from medical records to social media; NLP based automation will be critical to fully analyze the text and speech data efficiently (SAS, 2018). The automatic manipulation and processing of natural languages like, speech, text, etc. by software is considered as Natural Language Processing; If that used in resume analysis with the help of computers and software it is Natural Language Processing (NLP) Based Parsing (Brownlee, 2017).

Named Entity Recognition Based Parsing

The task of Named Entity Recognition and Classification can be described as the identification of named entities in computer readable text through annotation with categorization tags for required information extraction. Not only is Named Entity Recognition a subtask of information digging and extraction, but it also plays a vital role in reference resolution, other types of disambiguation, and meaning representation in other natural language processing applications. Semantic parsers, part of speech taggers, and thematic meaning representations could all be extended with this type of tagging to provide better results (Data Community, 2013). Named Entity Recognition is a process where an algorithm takes a string of text i.e., sentence or paragraph as input and identifies relevant nouns such as people, places, and organizations mentioned in the string into consideration. News and publishing houses produce large amounts of online content on day to day basis and to manage them correctly is very important to get the most use of each and every article. Named Entity Recognition can automatically scan entire articles and reveal which are the major people, organizations, and places discussed in them. Knowing the relevant tags for each article help in automatic categorization of the articles in defined and desired hierarchies and it enables smooth content discovery automatically and speedily (Gupta, 2017).

Co-reference Resolution Based Parsing

Co-reference resolution is the task of determining linguistic expressions that refer to the same real-world entity in natural language/s. In the sentences, like, "Have reviewed the electrocardiogram; It shows a wide QRS with a normal rhythm but no delta waves", the phrases "the electrocardiogram" and "It" refer to the same entity, i.e., the electrocardiogram. Co-reference consists of two linguistic expressions: antecedent and anaphor; the *anaphor* is the expression whose interpretation i.e., associating it with an either concrete or abstract real-world entity depends on that of the other expression. The *antecedent* is the linguistic expression on which an anaphor depends. "The electrocardiogram" is the antecedent, and "It" is the anaphor (Zheng, et. al, 2011). Co-reference is typical of anaphora realized by pronouns and non-pronominal definite noun phrases, but does not apply to varieties of anaphora that are not based on referring expressions, like verb anaphora. However, every noun phrase does not trigger co-reference. Bound anaphors which have as their antecedent quantifying noun phrases, like, as every man, most computational linguistics, nobody, etc. are another example where the anaphor and the antecedent do not co-refer (Mitkov, et. al., 2012). A parsing that uses the qualities of making such a distinction in digging out required set of information in a systematic manner out of a text, speech, etc. to provide a systematic output set in resume parsing is called as co-reference resolution based parsing.

Review of Literature

A system was proposed by the researchers, which allows the candidates to upload their resumes in flexible format. These resumes are then analyzed by the system developed by the researchers, indexed and stored in a specific format; it made the search process easy. The analyzing system worked on the algorithm used Natural Language Processing as sub domain of Artificial Intelligence. It read the resumes and understood the natural language/format created by the candidate and transformed it into a specific format. The so acquired knowledge was then stored in the knowledge base. The system acquired more information about candidates from their social profiles like, LinkedIn and Github and updated the knowledge base (Sadiq, et. al., 2016).

Researchers collected resumes from various websites like, naukri.com, LinkedIn, etc. having CV's uploaded by the applicants who seek for job. The uploaded CV's were in different formats like, doc and pdf format. Retrieving and analyzing data from pdf files is easier when compared to doc files, as word parsing is difficult and there are no free APIs (Application Program Interface) present. So, doc files were converted into pdf files first, and then pdf file were converted to text file using resume parser. Nouns are then extracted from the text file document by using Maxent Tagger. After conversion and extraction of resumes, researchers build the concept map which was done using the Hadoop framework based on MapReduce model. A qualitative assessment of resumes on the basis of different quality parameters using a simple text analytic based approach for a resume collection was described. The resume collection was assessed for two qualitative aspects, coverage and comprehensibility; and these ratings were transformed into a comprehensive quality rating. All the three parameters were collectively measured into a combined 1 to 5 rating scale for associating a quality metric for resumes. The qualitative evaluation results obtained through the algorithmic approach was congruent to and hence validated through the wisdom of crowds. Although researchers evaluated and combined two qualitative parameters for resume assessment in a systematic and thorough manner, some improvements and extensions were possible. The pdf parsing and section identification can also be improved. The standard reference documents could be augmented as well. Transforming the computed values to ranks has been the trickiest part. While for coverage, it was simpler; but in case of comprehensibility, it was a bit complex and tricky to transform computed values to 1 to 5 scale rating. Nevertheless, the algorithmic formulation had the possibility of being used in an annotation and recommendation system (Kudatarkar, Ramannavar and Sidnal, 2015).

An ontology-driven information extraction system called as Ontology Format Based Resume Parser (ORP) was proposed. The overall objective of the ORP system was based on a concept matching task and ontological rules for English and Turkish resumes that provided semantic analysis of data and parse related information such as experience, features, business and education information from a resume. The system contained various ontologies in its own Ontology Knowledge Base (OKB). Turkish and English clarifications were used for a better comprehension of the system mechanism for the case study section of the articles. The system had its own Semantic Matching Step (SMS) that was applied between two concepts, one from a resume and the other from related domain ontology in the OKB, to calculate a similar score. To conclude, the working system mechanism, the OKB, the matching steps, the transfer of plaintext resume into ontology form, the case studies, and the inference mechanism though the Semantic Web Rule Language (SWRL) rule base were discussed (Celik, 2013).

Researchers proposed a two-layer model for information extraction from resume documents in pdf format. This was done to take advantage of the layout and the content information of pdf documents. Various kinds of features of both content and layout were integrated. In the first layer, the resume documents were segmented into blocking by using heuristic rules. Then a well trained classification model was employed to classify each block segmented into blocking by using its heuristic rules. Afterwards, a well trained classification model was employed to classify each block into pre-defined categories. In the second layer, the detailed information extraction task was regarded as a sequence labeling problem. A Conditional Random Fields (CRF) was utilized to finish sequence labeling. The experimental results showed that the average F1 score (conveys the balance between the precision and the recall of the hierarchical extraction model) achieved 72.78%, which was 25 percent higher than the flat model. Besides, layout-based features were verified to be useful with a 22% improvement in average F1-score. The experiments demonstrated that the method could achieve high extraction accuracy with well adaptability to various document layouts (Chen, Gao and Tang, 2016).

Objective

To analyze and compare the contemporary resume parsing tools with HR perspective.

Rationale

Sorting of resume is very tedious time consuming job hence parsing becomes inevitable. This research paper focuses on the advancement and utility of predictive analytics for HR professionals via, NLP based resume parsing. NLP based resume parsing process is such an intelligent way that it could reach near human accuracy. This paper also flashes light on the advancement of NLP based resume parsing over Regular Expression/Rule Based Parsing.

Research Design

This research paper is experimental and case based action research. The features and utility of resumes parsers were analyzed, compared and concluded.

Hypothesis

There is no difference among the different resume parsers.

Limitations

Only trial versions / freeware were analyzed. Being into academic research work, the paid versions and demos were could not be arranged.

The Action Research

The Methodology

The methodology of the action research focused on understanding the various features of available resume parsers. The basis of analysis and comparison were decided as accuracy level, cost, ease of usage, extraction from social media, free trial availability, language, number of days of free trail available, number of features that can be parsed at a time, requirement of JD, speed of parsing and type of parser.

Actions

Literary meanings of different words were found out to describe resume parsing. Literature review was also done. Netsurfing found out following available resume parsing tools, a few names are: Adecco, Antwerp, Bullhorn, Candidatezip, Capterra, Crelate, CVlizer, Daxtra, HireAbility, iZito, Jobscience, Rapid Parser, rChilli, Resumefox, ResumeMill, Sovren, Talentlyft, Textkernel, Workable, Workato, Workopolis, Zapier, Zoho, etc.

Out of these, HireAbility, Jobscan, Rapid Parser, ResumeMill, Sovern, Talentlyft, Workable, Zoho were the ones with free trials. Four resume parsing tools ie., Zoho, Workable, Jobscan and Rapid Parser were selected as sample for testing to shed light on their qualities. Job Description (JD) of Senior Data Scientist was framed and used. 75 real resumes of various profiles were collected as sample from different sources, out of which 13 resumes are of pdf format and 62 resumes are in MS Word format. An analysis of the four parsers with considerable free trial account was under taken by uploading the real resumes on each of the parser and job descriptions individually in each of the four cases.

The Analysis and Outcome

The Analysis

Case Study 1: Jobscan (https://www.jobscan.co/)

It is termed as learning center; it helps candidates in getting desired job. It assists in preparing proper resume, cover letter, job interview, provides tips for applying on social job sites like LinkedIn, etc. Resume parsing is one of its features for assessment of resume against the job opening.

Features

Its features are: It matches resumes against the JD and provide results in percentage; Parses under categories, skills and keywords, job title, education, industry depth, hard skills, soft skills, etc.

Limitations

Its limitations are: It does not parse the contact information like phone number, email ID, etc.; It does not link the social network sites of candidates; It does not accept resumes without JD.

Case Study 2: Rapid Parser
(https://www.rapidparser.com)

It is basically a resume parser, parsing is its main functionality. It provides paid as well as free trail account.

Features

Its features are: It is fast as compared to others, it can parse around 100,000 resumes in a day; It supports maximum languages like English, Romanian, Slovenian, Czech, Turkish, French and Italian; It parses under categories like personal data, languages known, education, and employment history; JD is not required for parsing resumes.

Limitations

Its limitations are: Parsing carry various errors, the sample resumes which are parsed via Rapid Parser got incorrect information is education history and gender; Cannot download and save the parsed resume; The format of parsing is fixed, cannot personalize as per the need; It does not link the social network sites of candidates.

Case Study 3: Workable (https://www.workable.com)

Workable is also an ATS, resume parsing is an additional feature of workable. They do provide paid as well as free trial versions. Proper job application form is to be filled for creating JD.

Features

Its features are: Parser extracts the social media profiles of the candidate, along with the profile photo from the social site; Parses on three criteria i.e., phone number, email ID and social media profiles; It can send email to candidates via workable; Evaluation and comments about candidate can be added.

Limitations

Its limitations are: Parsing is done on only three criteria; There is no option to personalize parsing fields; Resumes cannot be uploaded without JD; No filtering of resumes as per the JD; Cannot download and save the parsed resume.

Case Study 4: Zoho (https://www.zoho.com)

Zoho is basically an ATS (Applicant Tracking System), data can be migrated from any other ATS to Zoho without any hassle. They provide free trail as well as purchased version. The free trail was used in this research; few information of JD has to be mentioned manually, can browse and import the job summary.

Features

Its features found are: Recruiters can be assigned against the JD, who is taking care of the said job opening; Interviews can be scheduled via Zoho; It can send SMS and emails to the candidates; The JD can be published via free job boards; Candidates can be associated against the JDs; Parsing feature extracts the Social Media profile of candidates which helps in crosschecking the candidates; Parsing is on criteria like name, mobile number, email ID, job title, employer, experience, qualification, etc.

Limitations

Its Limitations are: No filtering of resumes as per the JD; Cannot download and save the parsed resume; The format of parsing is fixed, cannot personalize as per the need; Cannot extract photograph of candidate from social media sites.

Outcome

S. No.	Name	Cost	Free Trail	Type	Type of Parsing	Requirement of JD
1.	Jobscan	$89.95 USD every 3 months after trial, $49.95 USD every month	30 Days	Resume Optimizer	Named Entity Recognition	Required
2.	Rapid Parser	500 Credits - $ 20, 1000 Credits - $35, 5000 Credits - $ 150, 10,000 Credits - $250 for 90 Days, 50% Discount on first order	Freely available but without saving option	Resume Parser	Regex	Not Required
3.	Workable	$2,500/year for 10 job slots and $4,000/year for 20 job slots	15 Days	ATS	Regex	Required
4.	Zoho	Rs.1350/- per month and Rs. 2700/- Annual	15 Days	ATS	Regex	Can upload but not must

S. No.	Name	Number of Features that can be Parsed at a Time	Extraction from Social Media	Ease of Usage	Accuracy Level	Speed of Parsing	Language
1.	Jobscan	7	No	Easy	At Par	High	English
2.	Rapid Parser	4	No	Easiest	Carries Error	Highest	Seven Languages
3.	Workable	3	Yes	Easy	At Par	High	English
4.	Zoho	7	Yes	Easy	At Par	High	English

Among the above mentioned parsers, Zoho seems to be the better one as it carries maximum required features, cost of purchase, 15 days free trail, extraction from social media, ease of usage, better accuracy. Jobscan is recommended in terms of matching with JD, It matches resumes against the JD and provide results in percentage; Parses under categories like, skills and keywords, job title, education, industry depth, hard skills, soft skills, etc., it also has considerable accuracy and ease of usage. Rapid Parser is better in terms of speed and functions in multiple languages.

Each one of them has some nice features as well as limitations which are mentioned in their description above. None of them are based on AI / NLP.

Discussion

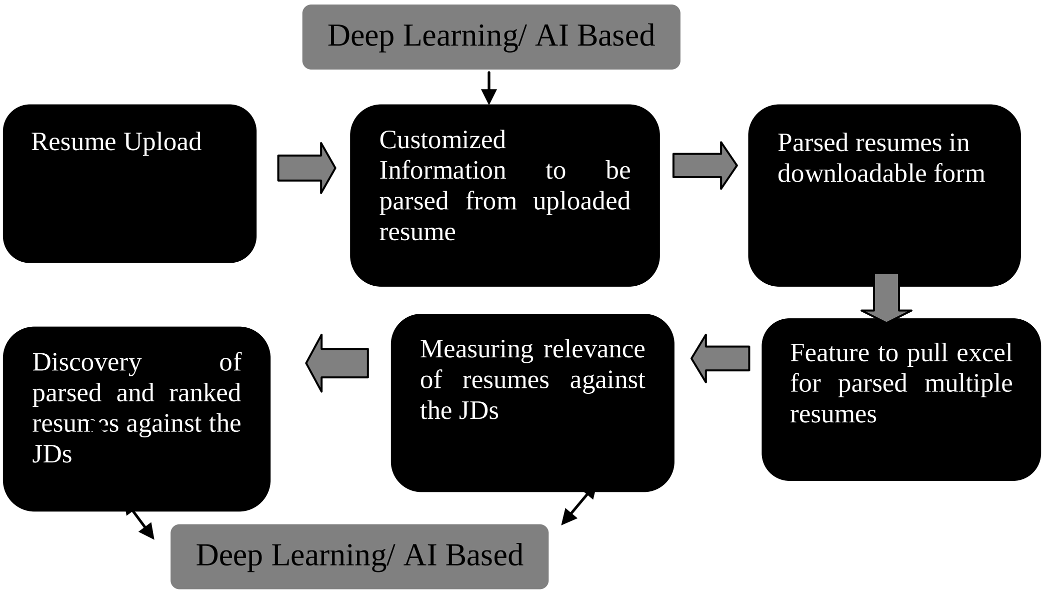
The resumes used were in PDF and MS Word format and the parsers used in the research are comfortable with both and cannot find any difference in parsing features between PDF and MS Word format. The main concern of resume parsers is the static features of regex and rule based parsing, they generally look for the set fields in resumes and they cannot move up to near human accuracy which is possible with NLP based parsing. They only process information which is set as a rule irrespective of the text which is mentioned in that field. There are innumerable resume formats available moreover each and every individual resume is different in one or the other features, one set rule for parsing may not work in each and every case; the need of NLP is to identify the required characteristics/fields irrespective of their place and format.

Findings/Results/Conclusion

The study of the above parsers provides their utility and limitations. Resume parsers definitely work as helping hand for recruiters. With the advancement of technology the resume parsers should also get more advanced to provide better results. They should provide comparative and valuation based results. Automated connectivity with social sites will add value in the work of resume parsing. Resume parsers should also get automated connectivity with professional sites to get more technical, professional and experience details about the candidates. Resume parser organizations should mention about the type of parsing they are using. Parsing should be a part of ATS/HRDBMS. Parsers should provide output in auto set format for all resumes.

Suggestions

On the basis of HR professionals/ recruiters being interviewed about resume parsers and as per the discussion and feedback from HR representatives the need of smart NLP based resume parsing system has been highlighted. A smart and ideal resume parser should carry the below mentioned features modeled as smart resume parsing tool:


Deep Learning/ AI Based
Resume Upload
Customized Information to be parsed from uploaded resume
Parsed resumes in downloadable form
Feature to pull excel for parsed multiple resumes
Measuring relevance of resumes against the JDs
Discovery of parsed and ranked resumes against the JDs
Deep Learning/ AI Based

The above smart model has advancements through deep learning/Artificial Intelligence (AI) based NLP. The system has to learn all the possible fields which a resume can have and with the help of AI it can provide the customized parsing fields as per requirement. For ranking of resumes against JDs the system has to learn all the possible fields present in any JD and then with AI it can match the fields of resume and JD.

Expected Advanced Features

Parsing output in excel format is the expected feature. Automated voice interviews should lead to speech analytics. Precise report should be forwarded by system to the concerned department for review/short listing for interview.

Implications

Resume parsing is the need of hour for HR recruiters and is a must for the up to date real time organizations. Advanced NLP based resume parsers ease out the recruitment process and ultimately save time of HR professionals for other productive activities. The customized resume parsing is of great help for recruiters in looking for only the required characteristics in any resume. The resume database created with the help of resume parser can help in future job openings. The HR professionals can provide valid justifications for short listed candidates on the basis of a set of resumes. NLP based parsers removes all sort of biases in resume screening process.

References

1. *Brownlee, Jason (2017). What is Natural Language Processing? Retrieved on June 25th, 2018 at 4:30 PM from https://machinelearningmastery.com/natural-language-processing/*
2. *Celik, Duygu (2013). Towards a Semantic-based Information Extraction System for Matching Resumes to Job Openings. Turkish Journal of Electrical Engineering & Computer Sciences, 24, 141-159. Retrieved on June 27th, 2018 at 6:00 PM from http://dergipark.gov.tr/download/article-file/429815*
3. *Chen, Jiaze; Gao, Liangcal and; Tang, Zhi (2016). Information Extraction from Resume Documents in PDF Format. Society for Imaging Science and Technology. Retrieved on June 28th, 2018 at 4:30 PM from https://www.ingentaconnect.com/contentone/ist/ei/2016/00002016/00000017/art00013?crawler=true*
4. *Computer Hope (2017). "Regex". Retrieved on June 19th, 2018 at 6:15 PM from https://www.computerhope.com/jargon/r/regex.htm*
5. *Data Community (2013). An Introduction to Named Entity Recognition in Natural Language Processing: Part 1. Retrieved on June 25th, 2018 at 6:00 PM from http://www.datacommunitydc.org/blog/2013/04/a-survey-of-stochastic-and-gazetteer-based-approaches-for-named-entity-recognition*
6. *Daxtra. (2016). What is CV/Resume Parsing? Retrieved on June 19th, 2018 at 5:00 PM from http://www.daxtra.com/2016/10/18/what-is-cvresume-parsing/*
7. *Gupta, Shashank (2017). Named Entity Recognition: Applications and Use Cases. Retrieved on June 25th, 2018 at 5:30 PM from https://towardsdatascience.com/named-entity-recognition-applications-and-use-cases-acdbf57d595e*
8. *Kudatarkar, Vinaya R.; Ramannavar, Manjula and; Sidnal, Nandini S. (2015). An Unstructured Text Analytics: Approach for Qualitative Evaluation of Resumes. International Journal of Innovative Research in Advanced Engineering, 8(2), 64-71. Retrieved on June 27th, 2018 at 5:30 PM from http://www.ijirae.com/volumes/Vol2/iss8/10.AUAE10097.pdf*
9. *Mitkov, Ruslan; Evans, Richard; Orasan, Constantin; Dornescu, Iustin and; Rios, Miguel (2012). Coreference Resolution: To What Extent Does It Help NLP Applications? Lecture Notes in Computer Science Book Series, United Kingdom, 7499 (16-27). Retrieved on June 25th, 2018 at 7:00 PM from https://link.springer.com/chapter/10.1007/978-3-642-32790-2_2*
10. *Rchilli. (2016). What is Resume Parsing? How CV Parser Works? Retrieved on June 19th, 2018 at 5:30 PM from https://www.rchilli.com/blog/resume-parsing-101/*
11. *Sadiq, Sayed Zainul Abideen Mohammad, Ayub, Afzal Juneja, Narsayya, Gunduka Rakesh, Ayyas, Momin Adnan and Khan, Tabrez Mohammad Tahir (2016). Intelligent Hiring with Resume Parser and Ranking using Natural Language Processing and Machine Learning. International Journal of Innovative Research in Computer and Communication Engineering, 4(4), 7437-7444. Retrieved on June 26th, 2018 at 7:00 PM from http://www.ijircce.com/upload/2016/april/218_Intelligent.pdf*
12. *SAS (2018). Natural Language Processing: What It is and Why It Matters. Retrieved on June 25th, 2018 at 5:30 PM from https://www.sas.com/en_us/insights/analytics/what-is-natural-language-processing-nlp.html*
13. *TechTarget (2006). Regular Expression (Regex). Retrieved on June 19th, 2018 at 6:00 PM from https://searchsoftwarequality.techtarget.com/definition/regular-expression*
14. *Zheng, Jiaping; Chapman, Wendy W.; Crowley, Rebecca S. and; Savova, Guergana K. (2011). Coreference Resolution: A Review of General Methodologies and Applications in the Clinical Domain. NIH Public Access, J Biomed Inform, 44(6), 1113-1122. Retrieved on June 25th, 2018 at 6:30 PM from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3226856/pdf/nihms-318898.pdf*