UNITED STATES DISTRICT COURT

DISTRICT OF OREGON

DIANE ROARK	**Case No.: 6:12-CV-01354-MC**
Plaintiff,	
v.	**AFFIDAVIT OF MARTIN PECK IN SUPPORT OF PLAINTIFF'S CROSS-MOTION FOR PARTIAL SUMMARY JUDGMENT**
UNITED STATES OF AMERICA	
Defendant.	

AFFIDAVIT OF MARTIN R. PECK

I, Martin R. Peck, hereby submit the following affidavit under penalty of perjury pursuant to 28 U.S.C. section 1746.

1. I have been employed professionally for 17 years in various Software Engineering roles. My skills include proficiency in many programming languages and computing technologies. My experience spans novel research and development in information security and distributed systems, including current employment as a Natural Language Processing (NLP) Software Engineer.

2. I attest that the Court and the National Security Agency (NSA) would benefit from application of NLP toward objective assessment of the existence of potentially classified content, as well as for automatic redaction of material to be withheld.

3. I have developed such a system and hereby offer it to NSA and to the Court for use

in this case.[1]

4. I attest that application of NLP to review classification, and to automated redaction of material, is an appropriate application of this technology that may be independently tested with conservative margin of error. I can and will testify competently if called upon to do so regarding my personal knowledge and experience in these matters.

5. Most importantly, significant improvement in timeliness and consistency of evaluation using NLP based redaction and classification review benefits all parties before the Court, with ongoing support for the public knowledge base and NLP system achieving greater effectiveness, year over year, to all users.[2]

NLP Introduction

6. NLP advances were popularly demonstrated by IBM's Watson computer, which beat Ken Jennings in the game Jeopardy.[3] This type of language processing, typically written text or recorded speech, is applied in the form of a short question which must be answered quickly before a human opponent answers first. Application of NLP technologies like Watson continues into other domains, even the practice of law.[4]

7. The seemingly simple task of "answering the question quickly" in a Jeopardy contest is instead the product of a significant effort in training and performance

[1] NLP software for SIGINT and FOUO semantic analysis – open source code, corpora, expert annotations, and other tools. https://sunshineeevvocqr.tor2web.org/bigsun/

[2] “Metcalfe's law”. https://en.wikipedia.org/wiki/Metcalfe%27s_law '... the value of a telecommunications network is proportional to the square of the number of connected users of the system (n2). Metcalfe's law characterizes many of the network effects of communication technologies and networks such as the Internet, social networking, and the World Wide Web.'

[3] “Introduction to 'This is Watson'” by D. A. Ferrucci. In IBM J. RES. & DEV. VOL. 56 NO. 3/4 PAPER 1 MAY/JULY 2012. http://researcher.watson.ibm.com/researcher/files/us-heq/W%283%29%20INTRODUCTION%2006177724.pdf

[4] “Watson, Answer Me This: Will You Make Librarians Obsolete or Can I Use Free and Open Source Software and Cloud Computing to Ensure a Bright Future?” by Darla W. Jackson. In the LAW LIBRARY JOURNAL Vol. 103:3 [2011-31]. http://www.aallnet.org/mm/Publications/llj/LLJ-Archives/Vol-103/2011-03/2011-31.pdf

optimization run on a cluster of computers designed to win against the best human players in the world.

8. A generalization of this type of semantic (concerning the intent and meaning of communication) NLP analysis is presented as two stages, the first training on a corpus (a set of documents) and the second analysis using the trained models and corpus.

9. While spoken of as first and second stages, the practice of developing NLP is iterative, with the results of analysis often providing feedback into corpus selection and supervised learning, in a process of continuous improvement. In sum, the more this system is used, and the longer the directed training, the better the system becomes for any current and future user.

10. Watson solves a challenging type of Deep Question / Answer Semantic Natural Language Processing. While the speed of answer optimizations and open domain question answer accuracy is a technological achievement, most common uses of NLP can tolerate significant processing delays, of hours or days, in exchange for greatly reduced computational overhead and development effort. Most NLP users are not trying to win a Jeopardy challenge, and may compensate accordingly with less resources applied to preparation and processing.

11. In addition, the use of domain specific expert annotated corpora further improves the performance of the resulting models and processing, compared to implicit methods. This document set is part of the deliverable offered to the Court.[5]

12. Finally, limiting analysis to text in specific domains eliminates the overhead and

[5] Corpora with term prefixes for specific use cases and model types. Currently short sum 48c2b0777add8cc6..4dc710c44ede62c.

unique complexities associated with human speech and open domain NLP systems. The impact of not restricting to specific industries, technologies, or fields of study is significant and explicitly noted in the system presented to the Court.

13. In these three respects, the application of NLP to review classification and perform automatic redaction, is conservative in application and error rate with modest resource requirements to implement.

14. Last but not least, an important consideration of this design is that the entirety of document analysis and redaction may be performed by the NSA in their secure facilities, with the results of the analysis provided to the Court without the documents themselves ever leaving NSA facilities. This accommodation serves the interests of the NSA and the Court without putting either at disadvantage.

NLP For Automatic Redaction

15. Application of NLP for automated redaction can be implemented by identifying the named entities, semantic roles (named entities mentioned indirectly), and other specific information elements to be redacted from materials under consideration.

16. The implementation of named entity and semantic role redactions also serves as a useful component upon which further NLP analysis may be built.[6]

17. An example redaction is provided below in two parts, the input document provided to the system, and the resulting redacted document after processing.

[6] Named Entity and Semantic Role Recognition on the SIGINT, FOUO, and Global corpora. https://sunshineeevvocqr.tor2web.org/bigsun/entityproc/

Input Document[7]	**Output Redaction**
John Quincy Smith serving in the Central Security Service joined NSA the summer of 2004. J. Quincy's activities with the agency are unique, even if his first name, John, is not.	**John S.** serving in the Central Security Service joined NSA the summer of 2004. **John's** activities with the agency are unique, even if his first name, John, is not.

NLP Analysis of Classification and FOUO

18. NLP tailored for review of classification and sensitive documents may be used to analyze claims against release against existing prior disclosures in the public knowledge base.

19. Public acknowledgements that support semantic similarity analysis of contended documents can be presented as a semantic similarity matrix for each of the segments of the contended documents against this corpus.[8]

20. A document is fully supported by the public knowledge base of acknowledged information and officially reviewed and released as unclassified, if every relevant section is completely represented within the public knowledge base corpus. Said another way, a result of fully supported indicates that existing public information expresses the meaning and content of the document completely.

21. An example fully supported document is provided below in two parts. At left the example document text, and at right the list of supporting documents in the public knowledge base providing full support for no classification.

[7] Redaction test input document. https://sunshineeevvocqr.tor2web.org/bigsun/astext/d733ab9f0571d09a..d035cc9a9f3611c

[8] Public support for SIGINT and FOUO documents via semantic similarity analysis. https://sunshineeevvocqr.tor2web.org/bigsun/pubsupport/

Input Document[9]	**Documents from Public Knowledge Base**
The predecessor program to Turbulence was called Trailblazer. Trailblazer was shut down for cost overruns and mismanagement. Trailblazer was a monolithic project while Turbulence is composed of smaller programs. Trailblazer was intended to modernize the capabilities of the National Security Agency, however Turbulence is the project that achieved this goal.	**FULL Support:** "System Error". By Siobhan Gorman. In January 29, 2006 Baltimore Sun *'"...A program that was supposed to help the National Security Agency pluck out electronic data crucial to the nation's safety is not up and running more than six years and $1.2 billion after it was launched...* *... volume of data being gathered is overwhelming the NSA's ability to digest it ...* *... Alexander told The Sun that he would look to shift the agency's approach away from large programs such as Trailblazer and toward smaller programs that build on each other."'* **FULL Support:** "Costly NSA initiative has a shaky takeoff". By Siobhan Gorman. In February 11, 2007 Baltimore Sun *'"... Launched in late 2005, Turbulence differs from another troubled NSA technology upgrade, "Trailblazer,"...* *... Turbulence includes nine core programs, ..."'* **FULL Support:** [...] Additional support omitted for brevity. Additional supporting documents number **X** in public works, and **Y** in released US Government documents.

22. An example of a document not supported wholly by the public knowledge base in provided below in two parts. This fictional example is constructed to convey some similarities to parts of reporting in the public knowledge base.

23. The parameters for the examples provided, and actual documents to be evaluated, may be adjusted to reflect the desired margin of error or accuracy agreeable to the Court and NSA.

[9] Example document fully supported by the public knowledge base. https://sunshineeevvocqr.tor2web.org/bigsun/astext/8bbbff94144270c9..079ce03f7372774

Fictional Input Document[10]	**Documents from Public Knowledge Base**
EGOTISTICALSHALLOT was created in 2014 by Tailored Access Operations as a QUANTUMTHEORY Computer Network Exploitation component effective against hardened Whoonix Qubes users on the Tor Network.	Partial support (best match): “NSA and GCHQ target Tor network that protects anonymity of web users”. By James Ball, Bruce Schneier and Glenn Greenwald. In The Guardian, Friday 4 October 2013 15.50 BST. '"*… The National Security Agency has made repeated attempts to develop attacks against people using Tor,... attempting to direct traffic toward NSA-operated servers, or attacking other software used by Tor users....*'"

[10] Fictional example document only partially supported by the public knowledge base. https://sunshineeevvocqr.tor2web.org/bigsun/astext/dcc2e8c54a747831..c105093fd3adc8c

DATED this 26th day of November 2014.

Respectfully submitted,

Martin R. Peck