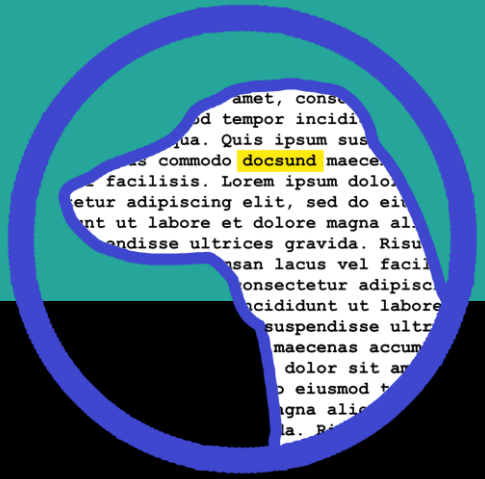
Docsund



Ryan Delgado, Andrew Carlson, Danielle O'Neil, Matthew Prout

August 5, 2019

You're a journalist and WikiLeaks just leaked a new dump of 1,000,000 documents...now what?

Existing process:

- Beat writers comb for keywords that are known to be relevant to them
  - i.e. I know I'm interested in writing about Uber; I'll search for different combinations of 'Travis' and 'Kalanick'
- CTRL+F
- At the same time, speed and accuracy are imperative, as you would like to be first to break the story.

Problem Definition

Why do they need this product?

When looking through a large document or email dump, investigative journalists and other researchers often have the following challenges:

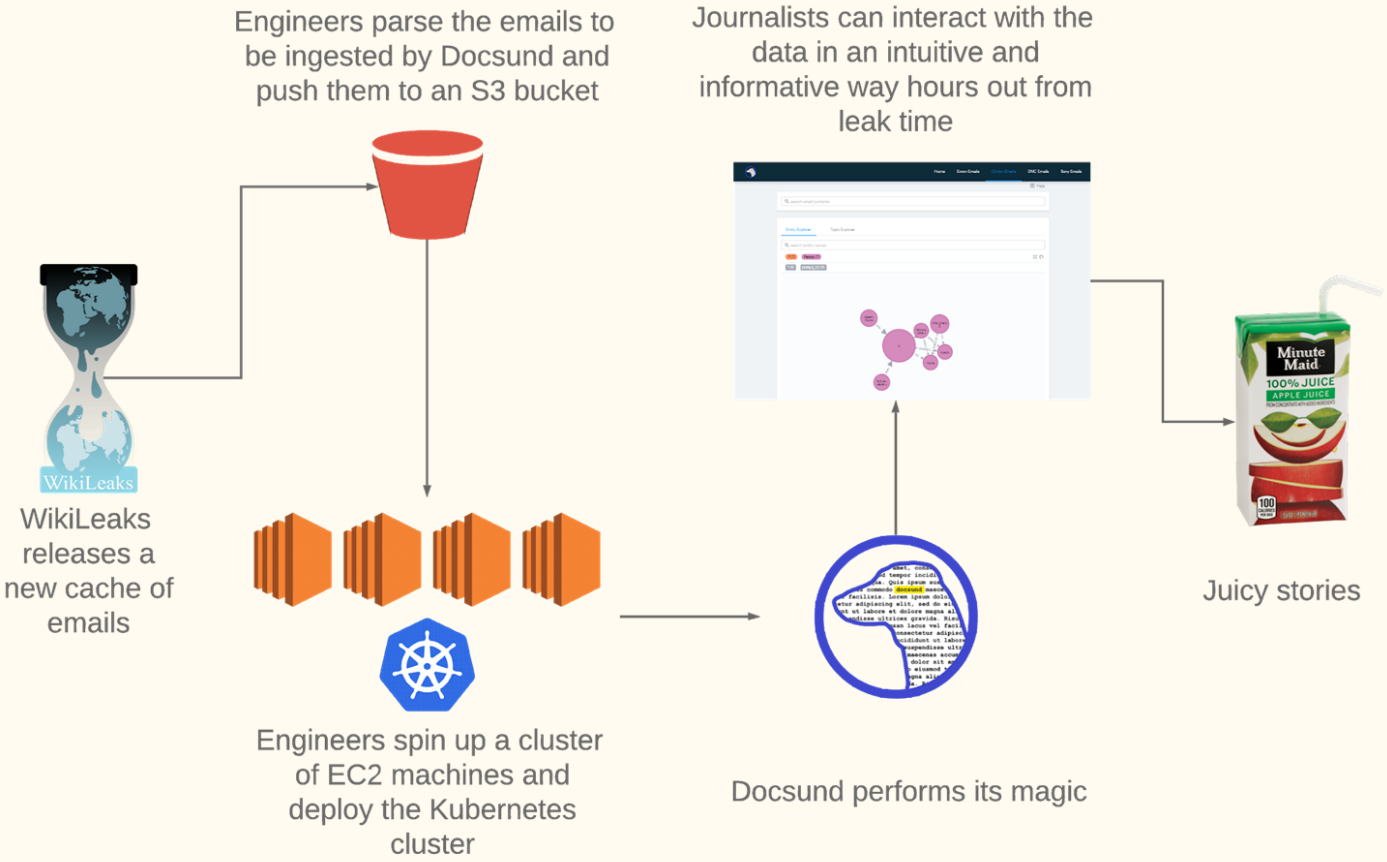
Unstructured Data	1. The documents are unstructured with no logical starting point.
	2. Often there is no way to see what topics exist in the documents.
	3. It is difficult to get a cohesive picture of the relationships between entities.
Large Size of Corpus	4. The user often needs to find insights quickly. For instance, investigative journalists want to be the first to break a story. But because the corpus of data is so large, traditional search methods are slow and ineffective. In this case the user relies on tedious guesswork.
Pronouns / Alternate Spellings	5. Traditional search methods will miss entities being searched for due to the use of pronouns and alternate spellings.

There isn’t currently a tool that applies NLP to large document dumps.

- In the internet age, sites like WikiLeaks are making large data dumps increasingly common
- Existing tools stop at visualizing relationships
  - Even this is unsatisfactory when context is key
- Quickly sifting through these large caches of documents is an issue journalists are going to face more often

Demo

Use Case Lifecycle



Value Proposition

The findings shown in the demo above were found organically in a matter of hours. We also found corresponding articles in the news surrounding these findings:

Defamer [Leaked: The Nightmare Email Drama Behind Sony's Steve Jobs Disaster]

Leaked on November 24, 2014, that article published on December 9 (15 days)

Los Angeles Times [Sony wrestles with how to market 'Interview' amid geopolitics, scandal]

Leaked on November 24, 2014, published on December 12 (18 days)

Stakeholder: Eric Newcomer

Throughout the project we worked with our primary stakeholder, Eric Newcomer, to define the use cases, features, and get his feedback on the design. Eric is a business reporter for Bloomberg, where he covers Tech IPOs.

- **First Meeting:** Learned about use cases. Reporters use text editors to search for terms they think may be relevant in the email dump. They are looking for names, financial numbers, and who works at the company. Reporters are seeking to answer the “who” and “what”.
- **Second Meeting:** Showed Eric the prototype we developed and got his feedback on the approach we are taking.
- **Third Meeting:** In this meeting we let Eric test the more finished product on his own and listened to his feedback. This meeting generated a long list of issues, many of which we were able to fix.
    - “Generally, this is the right idea” - Eric
- **Fourth Meeting:** In this meeting we let Eric test the product after making changes from the last meeting.
    - “Over all, this is in a place where someone can take advantage of it - to find new things" - Eric
    - "[It] puts data out there in a new lens" - Eric

We succeed when journalists succeed.

In a time where journalism and free press are more important than ever, getting tools like this to journalists:

- Increases their credibility
- Enhances their ability to dive deeper in early breaking stories
- Keeps journalism relevant in the internet age

Questions?

Website: http://docsund.info/

Project Attribution, by Component

- **Entity Explorer** - Developed by Andy & Danielle
- **Topic Explorer** - Matt, Andy, and Ryan
- **Search Bar** - Matt and Andy
- **Feature extraction pipeline** - Everybody
- **Infrastructure** - Andy
- **Data acquisition** - Ryan