
Text Analysis for eRulemaking

Jamie Callan

Carnegie Mellon University

callan@cs.cmu.edu

Human Languages and Computers

- **Public comments are expressed in human languages**

- “The term organic should never include genetic engineering or irradiation. Please change your rules to make this absolutely clear. These are totally unacceptable practices.”

Comment w0000050, received 12/16/1997, USDA National Organic Program, Docket Number: TMD-94-00-2

- **Computers don't understand human languages the way we do**

- How can they be useful?

Language Technologies

- **Computers can sometimes do useful things with human language ...even though they don't understand the language like we do**
 - Example: Web search engines
 - Example: Simple voice recognition on cell phones
 - Example: Monitoring newswires for “interesting” events
 - Example: Creating searchable job databases from online job postings
- **The systems that do these tasks are not perfect**
 - But accuracy is often comparable to ordinary people
 - They enable people to work with far more information
 - » E.g., carpenters and power saws

Introduction to Simple Statistical Language Technologies

- **Recognize words and phrases**
 - Handle morphology (e.g., plurals, past & present tense, etc)
 - Recognize parts of speech (e.g., nouns and verbs)
- **Count how often things occur**
 - In general
 - Or as compared to a reference model (e.g., general English)
- **Count how often two things occur together**
 - In general, or as compared to a reference model
- **Measure similarity of documents / paragraphs / sentences**
 - Based on the words they have in common
- **Sort things into categories (based on similarity)**
- **Recognize simple syntactic relationships and patterns**

(Simple) Text Analysis for eRulemaking

- **Search**
- **Information filtering (Selective Dissemination of Information)**
- **Duplicate and near-duplicate detection**
- **Frequent concepts**
- **Clustering**
- **Classification**
- **Opinion mining**
- **Stakeholder identification**
- **Stakeholder relationships / organization**
- **Information extraction**

(Simple) Text Analysis for eRulemaking

- Search
- Information filtering (Selective Dissemination of Information)
- Duplicate and near-duplicate detection
- Frequent concepts
- Clustering
- Classification
- Opinion mining
- Stakeholder identification
- Stakeholder relationships / organization
- Information extraction

Duplicate Detection

- **Problem: Many comments are form letters or edited form letters**
 - E.g., created by organized interest groups and lobbies
- **Solution: Duplicate and near-duplicate detection algorithms**
 - Generate summary counts
 - Identify the reference copy
 - Summarize differences from reference copy
- **Near-duplicate detection**
 - Use clustering to identify similar documents
 - Identify near-duplicates using document “fingerprints”
 - » Sequences of words that match in each document



File

Tools

Help

Basic Statistics

Dupli

Frequ

Stake

schakel, jonathan (04/28/1998)

Docket:

TMD-94-00-2 Topic:

National List Section:

205.22 (Subpart B - The Use of Active Synthetic Substances, Non-synthetic Substances, Non-Agricultural (non-organic) Substances and Non-organically Produced Ingredients in Organic Farming and Handling Operations, Including the National List of Allowed and Prohibited Substances) **My** comment is in **regard** to Section 205.22 of the National Organic Production Rule proposed December 16, 1997. I understand that the USDA has added numerous items to this list that were not recommended by the National Organic Standards Board. I consider this is a violation of the Organic Food Production Act of 1990 Section 6516(d)(2).

I would like to see the following revisions to Section 205.22:

Act of 1990 Section 6516(d)(2).

I would like to see the following revisions to Section 205.22:

....

This**reference**

Near Duplicate Detection: Example

Original Document: 121

Cheryl Peppel
12101 Freeman Ave
Hawthorne, CA 90250

February 19, 2004

USDA Regulatory Analysis Development
PDD, APHIS, Station 3C71
4700 River Road, Unit 118
Riverdale, MD 20737-1238

Dear USDA Development:

These comments are about Docket No. 03-031-2 regarding the Environmental Impact Statement for genetically engineered crops and organisms.

The USDA has not done enough to safeguard the environment from the potential damage that could be caused by genetically engineered crops. Moreover, the USDA is engaged in policies that are benefiting the biotech industry but hurting the organic industry.

Evidence shows that organic corn is becoming contaminated by the cross-pollination from genetically engineered corn. Further, by allowing the growth of so many acres of crops that contain the Bt (bacillus

Modified Document: 71

John and Lucy Perko
1017 Grandview Ave.
Ojai, CA 93023

February 3, 2004

USDA Regulatory Analysis Development
PDD, APHIS, Station 3C71
4700 River Road, Unit 118
Riverdale, MD 20737-1238

Dear USDA Development:

These comments are about Docket No. 03-031-2 regarding the Environmental Impact Statement for genetically engineered crops and organisms.

The USDA has not done enough to safeguard the environment from the potential damage that could be caused by genetically engineered crops. Moreover, the USDA is engaged in policies that are benefiting the biotech industry but hurting the organic industry.

Evidence shows that organic corn is becoming contaminated by the cross-pollination from genetically engineered corn. Further, by allowing the growth of so many acres of crops that contain the Bt (bacillus

Near Duplicate Detection: Example

Dear USDA Development:

These comments are about Docket No. 03-031-2 regarding the **Environmental Impact Statement** for genetically engineered crops and organisms.

The USDA **has not done enough to safeguard the environment from the potential damage that could be caused by genetically engineered crops.** Moreover, the **USDA is engaged in policies that are benefiting the biotech industry but hurting the organic industry.**

Evidence shows that organic corn is becoming contaminated by the cross-pollination from genetically engineered corn. Further, by allowing the growth of so many acres of crops that contain the Bt (bacillus thuringiensis) toxin, insects are beginning to build up resistance to the Bt toxin due to overexposure. As a result, the effectiveness of Bt, when used as a spray in organic agriculture, is undermined.

In addition, recent research shows that a lot of "horizontal gene transfer" is taking place in ways that many genetic researchers earlier thought was impossible. And biologists have been shocked to discover plant species with differing numbers of chromosomes are hybridizing. Gene transfer has been shown to occur in open fields of canola, resulting in the creation of "superweeds" that have become herbicide-resistant.

Far more is still unknown about the science of genetic engineering than is known. For example, just a few years ago scientists decided to map the human genome and determine the genes that make up a human. They expected to find more than 100,000 genes. But scientists were shocked to discover that humans are made out of only about 30,000 genes. Scientists determined

Dear USDA Development:

I am writing on behalf of Oregon Physicians for Social Responsibility, representing over 850 Oregonians. These comments are about Docket No. 03-031-2 regarding the **EIS** for genetically engineered crops and organisms.

The USDA **should be protecting the public health and safety of the environment rather than supporting the biotech industry.** Genetically engineered foods especially those grown in open air situations have not had enough testing to declare them safe. To have the **entire European Union banning GMO's gives one pause.**

First, we know that cross pollination will occur and that even with buffer zones most genes are found further afield even if the farmers complied which they don't due to loss of acreage that can be planted. There may be some genetic changes that are ultimately useful such as the insertion of a thiamine producing gene in rice but the insertion of toxins to wipe out pests will only yield stronger, more genetically "smart" insects due to overexposure and genetic drift.

In addition, recent research shows that a lot of "horizontal gene transfer" is taking place in ways that many genetic researchers earlier thought was impossible. And biologists have been shocked to discover plant species with differing numbers of chromosomes are hybridizing. Gene transfer has been shown to occur in open fields of canola, resulting in the creation of "superweeds" that have become herbicide-resistant.

Public Comment: Example w0000050

received date = 12/16/1997

name = *****, *****

company =

state = GA

zip = 30295

country = United States

category = Consumer

topic = Applicability

section = 205.3

comment = The term organic should never include genetic engineering or irradiation. Please change your rules to make this absolutely clear. These are totally unacceptable practices.



File

Tools

Help

Basic Statistics

Frequent Concepts (Noun Phrases)

Phrase	Count	Phrase	Count
sewage sludge	5,590	docket #tmd	572
national organic standards board	4,044	genetic engineering	568
national list	2,855	term effects	564
food production	1,897	organic foods	554
food irradiation	1,187	federal register	535
food products	1,015	national organic production rule	524
food supply	945	dear ms	522
national organic program	909	united states	432
nosb recommendations	851	livestock feed	421
organic foods production act	837	livestock production	417
organic food production act	817	crop production	369
national organic standards	803	growth hormones	366
sewer sludge	759	factory farming	365
organic food	598	animal cannibalism	365
food industry	581	animal products	364

Public Comment: Example w0010435

received date = 04/03/1998

name = *****, *****

company =

state = NJ

zip = 07739

country = United States

category = Consumer

topic = General

comment = As a mother and consumer of organic products, I **strongly oppose**
the **proposed regulations**

My family purchases organic food so that we know what we are eating. I understand the current standards that have been self-imposed by the organic food industry. ...

eRulemaking: Opinion Mining

- **There is increased focus on detecting opinions in text**
 - Using basic text categorization techniques
- **The state-of-the-art is not very good yet**
 - E.g., in a test using Epinions.com data:
 - » Cars vs. digital cameras: 99% correct
 - » Ford vs. other auto manufacturers: 87% correct
 - » Recommend Ford vs. Don't recommend Ford: 70% correct
- **This may look discouraging, but perhaps it isn't**
 - Very simple techniques
 - Significant attention being paid to improving accuracy

Opinion Mining for eRulemaking: Next Steps

- **Identify topic**
 - E.g., “sewage sludge”
- **Identify text with an opinion**
 - E.g., “I do not want sewage sludge to be used on any ‘organic’ food that I eat.”
- **Classify the opinion**
 - E.g., Against

Public Comment: Example w0010435

received date = 04/03/1998

name = *****, *****

company =

state = NJ

zip = 07739

country = United States

category = Consumer

topic = General

comment = As a **mother** and consumer of organic products, I **strongly oppose** the proposed regulations.

My family purchases organic food so that we know what we are eating. I understand the current standards that have been self-imposed by the organic food industry. ...

Public Comment: Example w0006200

received date = 03/06/1998

name = *****, *****

company = Columbia College Chicagoo

state = IL

zip = 60626

country = United States

category = Consumer

topic = General

comment = Thank you for providing me with the opportunity to make these remarks concerning the Proposed Rule (docket number TMD-94-00-2).

I am a doctor of engineering, a writer and a consumer of organic products.

(1) I am concerned that the Proposed Rule does not take the previous usage of agricultural land into consideration in the certification process. This seems unrealistic, given the lingering character of many industrial pollutants.

Automatic Stakeholder Identification

- **A stakeholder is anyone who has an interest in a topic**
 - Regulators like to know which stakeholder groups are represented in the comments, and in what proportions
- **People often identify which stakeholder groups they belong to**
 - “As a family farmer, I believe....”
 - “As a long-time consumer of organic foods, I am horrified...”
- **Simple patterns can recognize common stakeholders**
 - They won’t get everything, but that’s okay
- **Very simple example rule**
 - “As a” + <token> + punctuation
 - » Token must be a noun, adjective, article, etc



File

Tools

Help

Basic Statistics

Duplicate Detection

Candidate Stakeholder	Count	Candidate Stakeholder	Count
consumer	697	scientist	13
consumer of organic foods	90	taxpayer	13
consumer of organic products	80	a consumer	12
mother	47	long time consumer of organic foods	10
result	38	american	9
consumer of organic produce	37	organic foods consumer	9
citizen	31	a parent	8
concerned consumer	25	biologist	8
organic gardener	21	consumer of organic goods	8
consumer of organic food	19	grower	8
person	17	long time consumer of organic products	8
concerned citizen	16	california resident	7
organic food consumer	14	consumer of organic food products	7
organic consumer	13	farmer	7



File

Tools

Help

Basic Statistics

Duplicate Detection

Candidate Stakeholders (Hierarchical Display)

- **consumer (195)**
 - organic
 - organic food/foods (109)
 - organic food products (7)
 - organic goods (8)
 - organic products (8)
 - organic produce (37)
 - : : : :
- **mother (61)**
 - homemaker / at home (3)
 - concerned (2)
 - working (5)
 - new (2)
 - of a baby/child/... (3)
- **<profession> (82)**
 - scientist (13)
 - biologist (8)
 - farmer (7)
 - grower (8)
 - organic gardener (21)
 - : : :
- **citizen (58)**
 - american (9)
 - taxpayer (13)
 - concerned (16)
 - : : :

Linking Comments to the Proposed Regulations

- **Rule writers might find it helpful to have comments organized by the rule sections on which they comment**
- **It's relatively easy to link comments that explicitly mention a section number**
 - Very easy if the section is mentioned in a fill-in-the-blank field
 - Not too hard if the section is mentioned in the text
- **It is possible to link comments that use key vocabulary associated with a section**
 - Results are somewhat less accurate...

Attracting Research Attention

How do new technologies get developed?

- **Show money on research institutions**
 - My email address is callan@cs.cmu.edu ☺
- **Provide interesting problems and data to researchers**
 - Academic researchers are starved for “real world” data
 - Academics like to find out where their techniques work & don’t
 - » Sometimes we learn more from failures than successes
- **Examples:**
 - Reuters
 - NIST’s TREC conference: Wall St. Journal, NY Times, ...
 - Library of Congress Digital Memories and THOMAS
 - : : : : :



Sponsored by the NSF Digital Government Program

- USDA National Organic Program
- USDA importation of solid wood packing material standards
- USDA environmental impact statement for genetically engineered crops
- US EPA Clean Water Act revisions
- USDOT hours-of-service (HOS) regulations
- USDOT Corporate Average Fuel Economy (CAFE) standards
- USDOT light sport aircraft
- USDOT filing procedures for OST docket
- USDOT commercial air tours

More datasets to come...

<http://www.cs.cmu.edu/~callan/Data/>

Language Technologies for eRulemaking: Elements of an Analyst's Workbench

Language technologies are power tools for humans

- Full-text search
 - Identify duplicates, near duplicates, and common variations
 - Identify common concepts
 - A starting point for drill-down activities
 - What co-occurs with common concepts
 - Route comments to area experts
 - Identify stakeholders
 - Are they the ones that were expected?
 - Link comments to sections of the proposed regulation
- : : : :

eRulemaking: State-of-the-Art

Agencies are struggling with how to integrate information technology (IT) into the rulemaking process

- **Basic IT issues**

- How to collect, provide access, preserve
- The 1997 National Organic comments are essentially inaccessible

- **Language technologies**

- There is a great opportunity to provide better analysis

- **Policy issues**

- Does eRulemaking lead to better rules?
- What are the consequences of eRulemaking?