



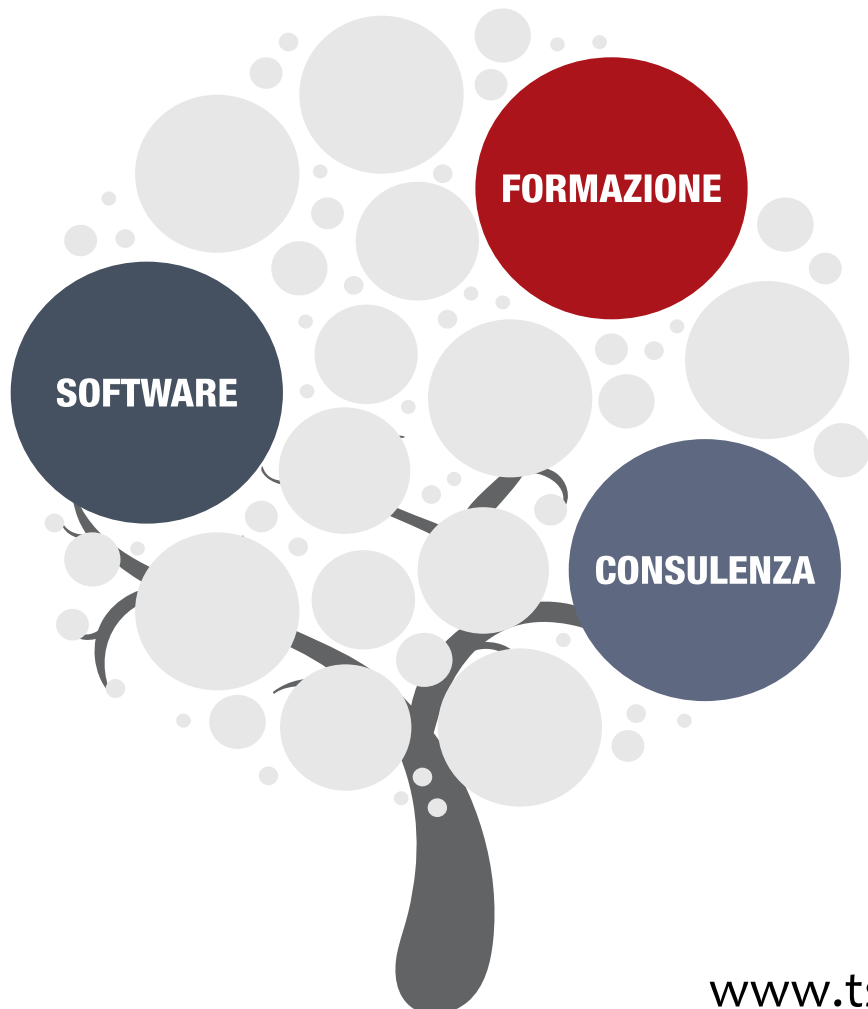
## QUANTITATIVE AND QUALITATIVE ANALYSIS

Normand Péladeau

President

Provalis Research Corp.

XII Convegno Italiano degli Utenti di Stata • Firenze, 12 Novembre 2015



[www.tstat.it](http://www.tstat.it)





Normand Péladeau  
President  
Provalis Research Corp.  
peladeau@provalisresearch.com

## Provalis Research

Some commercial clients:



## Provalis Research

Some governmental and NGO clients:

The slide displays a collection of logos for governmental and NGO clients. The logos are arranged in a grid-like fashion. Key clients include the European Commission, NASA, Statistique Canada, INRA, NHS, CERN, RAND Corporation, MITRE, United States Postal Service, Ministry of Defence, Federal Reserve System, CDC, U.S. Air Force, Johns Hopkins Medicine, Institut Pasteur, World Vision, and Australian Government Department of Human Services.

## Provalis Research

Some Academic Clients:

The slide displays a collection of logos for academic clients. The logos are arranged in a grid-like fashion. Key clients include Harvard University, HEC Paris, LSE, The Chinese University of Hong Kong, Université de Montréal, University of Cambridge, Ecole Hôtelière de Lausanne, MIT, Universidad de Chile, University of Copenhagen, Université Paris Descartes, The University of Melbourne, ESC Rennes, Maastricht University, Sapienza Università di Roma, Universität Hamburg, Seoul National University, Universidad Autónoma de Madrid, and University of Helsinki.

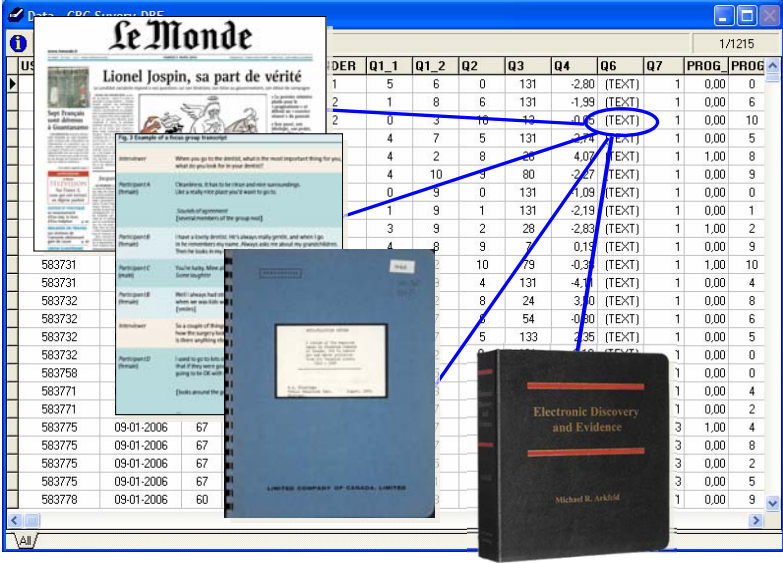
## Our Products

**1989**



**Statistical Analysis  
& Bootstrapping**

## Our Products



DER	Q1_1	Q1_2	Q2	Q3	Q4	Q6	Q7	PROG	PROG
1	5	6	0	131	-2.80	(TEXT)	1	0.00	0
2	1	8	6	131	-1.99	(TEXT)	1	0.00	6
2	0	3	10	12	0.00	(TEXT)	1	0.00	10
4	7	5	131	12	4.07	(TEXT)	1	0.00	5
4	2	8	26	4.07	(TEXT)	1	1.00	8	
4	10	3	80	-2.87	(TEXT)	1	0.00	9	
0	9	0	131	-1.09	(TEXT)	1	0.00	0	
1	9	1	131	-2.19	(TEXT)	1	0.00	1	
3	9	2	28	-2.83	(TEXT)	1	1.00	2	
4	8	9	7	0.19	(TEXT)	1	0.00	9	
		10	79	-0.31	(TEXT)	1	1.00	10	
		4	131	-4.11	(TEXT)	1	0.00	4	
		8	24	3.10	(TEXT)	1	0.00	8	
			54	-0.80	(TEXT)	1	0.00	6	
			5	133	2.35	(TEXT)	1	0.00	5
							1	0.00	0
							1	0.00	0
							1	0.00	4
							3	0.00	8
							3	0.00	2
							3	0.00	5
							1	0.00	9

## Our Products

1989



**Statistical Analysis  
& Bootstrapping**

1998



**Content Analysis  
& Text Mining**

2004



**Qualitative Analysis  
& Mixed Methods**

## WordStat for Stata

June  
2013



**Stata 13**

**Content Analysis  
& Text Mining**

REVIEW	ANNO	Variables
1	2013	REVIEW
2	2013	REVIEW
3	2013	REVIEW
4	2013	REVIEW
5	2013	REVIEW
6	2013	REVIEW
7	2013	REVIEW
8	2013	REVIEW
9	2013	REVIEW
10	2013	REVIEW
11	2013	REVIEW
12	2013	REVIEW
13	2013	REVIEW
14	2013	REVIEW
15	2013	REVIEW
16	2013	REVIEW
17	2013	REVIEW
18	2013	REVIEW
19	2013	REVIEW
20	2013	REVIEW
21	2013	REVIEW
22	2013	REVIEW
23	2013	REVIEW
24	2013	REVIEW



## Text Analytics Challenge

### **THREE MAJOR OBSTACLES**

1) Very large number of word forms

## Challenge #1 – Quantity

**38,996 comments about hotels**

- 2,1 millions words (tokens)
- 20,116 word forms (types)

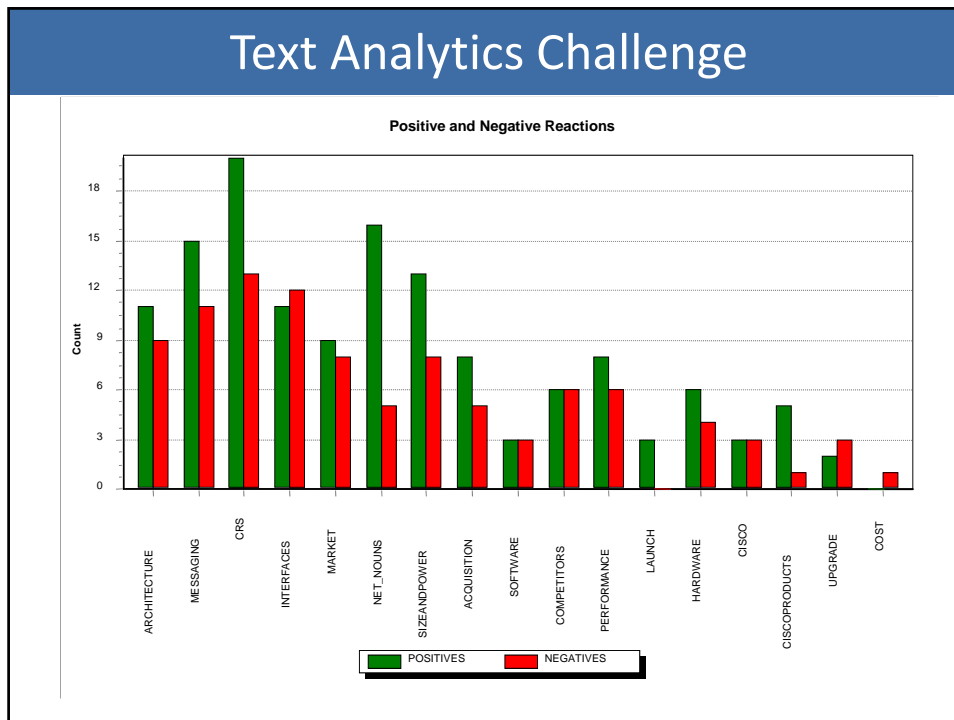
**1,8 million course evaluations**

- 35 millions words (tokens)
- 78,159 word forms (types)





## Text Analytics Challenge



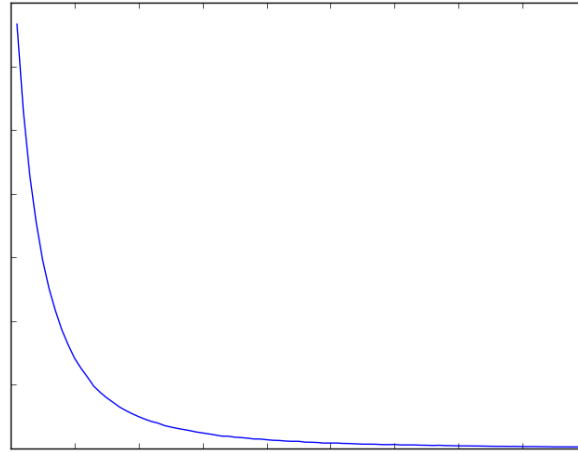
## Challenge #1 – Solutions

### Statistical tools

- Frequency selection
- Data reduction techniques (HCA, PCA, FA)
- Exploratory data analysis (ex. CA).
- Machine Learning

## The Statistics of Text

### Distribution of words: Zipf distribution



## The Statistics of Text

38,988 comments about hotels

2.1 M words (20,114 different words)

MOST FREQUENT FORMS	PERCENTAGE OF FORMS	PERCENTAGE OF WORDS
49	0.24%	50%
300	1.5%	76%
500	2.5%	83%
1000	5.0%	90%

## Challenge #1 – Solutions

### Statistical tools

- Frequency selection
- Data reduction techniques (HCA, PCA, FA)
- Exploratory data analysis (ex. CA).
- Machine Learning

### Natural language processing (NLP) tools

- Stopword list
- Stemming
- Lemmatization



## Text Mining Approach

### **PROS**

- Very fast
- Very little efforts
- Inductive

### **CONS**

- Comparison of results
- Does not quantify accurately
- Insensitive to low frequency events
- Inductive

## Text Analytics Challenge

### **THREE MAJOR OBSTACLES**

- 1) Very large number of word forms

## Text Analytics Challenge

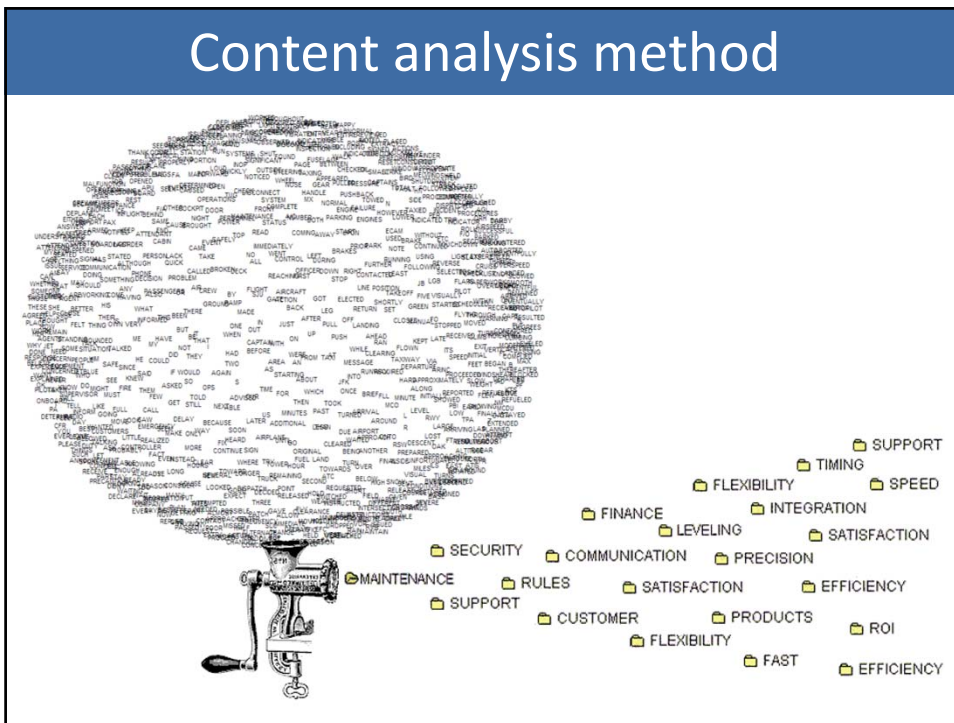
### **THREE MAJOR OBSTACLES**

- 1) Very large number of word forms

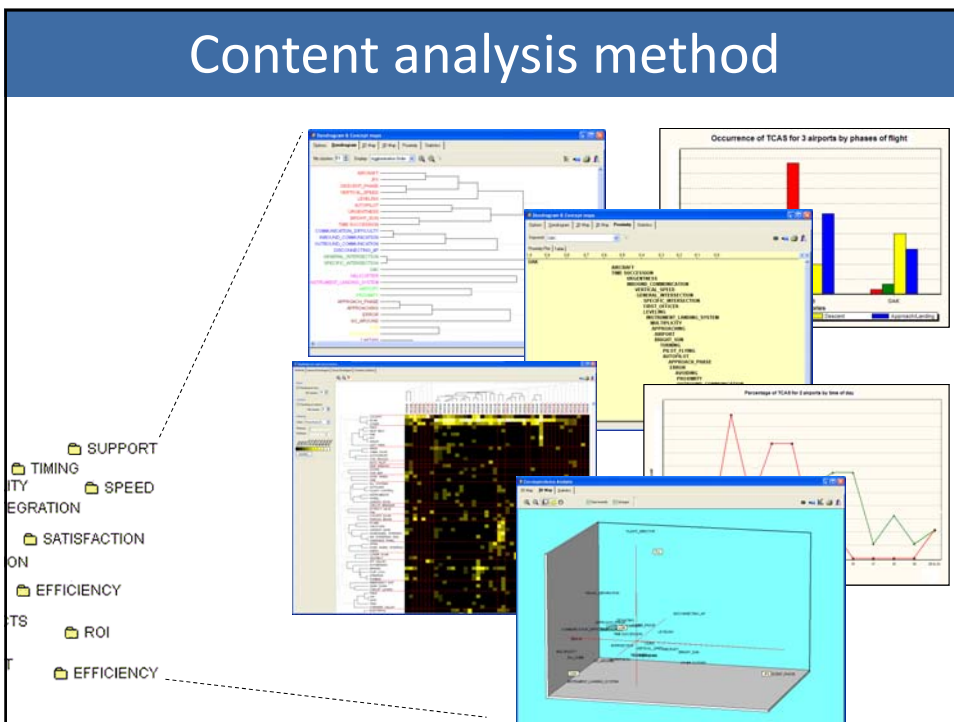
- 2) Polymorphy of language

One idea → multiple forms

# Content analysis method



# Content analysis method



## Content analysis method

- ☐ SENSE\_OF\_HUMOR
- ☐ SOCIALISATION
- ☐ CONFLICT\_RESOLUTION
- ☐ TEAM\_WORK
- ☐ ORGANIZATION\_SKILLS
  - ORGANIZATION\_SKILLS (1)
  - ORGANIZATIONAL\_SKILLS (1)
  - TIME\_MANAGEMENT (1)
  - COMPLETE\_A\_TASK (1)
- ☐ PROBLEM\_SOLVING
  - PROBLEM\_SOLVE (1)
  - PROBLEM\_SOLVE\_SKILL (1)
  - PROBLEM\_SOLVING (1)
  - PROBLEM\_SOLVING\_SKILLS (1)
  - SOLVE\_A\_PROBLEM (1)
  - SOLVE\_PROBLEMS (1)
  - SOLVING\_PROBLEMS (1)
- ☐ LEADERSHIP
  - LEAD\_A\_GROUP (1)
  - LEAD\_A\_TEAM (1)
  - LEAD\_GROUPS (1)
  - LEAD\_PEOPLE (1)
  - LEADERSHIP\_SKILLS (1)
  - LEADING\_A\_GROUP (1)
  - LEADING\_A\_LARGE (3)
  - LEADING\_A\_TEAM (1)
  - LEADING\_GROUPS (1)
  - LEADING\_TEAMS (1)
  - TAKING\_CHARGE (1)
  - TAKING\_COMMAND (1)
  - TAKING\_THE\_LEAD (1)
  - GROUP\_LEADERSHIP (1)
  - TEAM\_BUILDING (1)
  - TEAM\_BUILDING\_SKILLS (1)
  - GUILD\_LEADER (1)
  - GROUP\_LEADER (1)
  - GOOD\_LEADER (1)
  - TEAM\_LEADER (1)
- ☐ CREATIVITY
  - CREATIVE\_OUTLET (1)
  - CREATIVE\_THINKING (1)
  - CREATIVE\_SIDE (1)
  - CREATIVE\_WRITING (1)
  - CREATIVE\_PROCESS (1)
  - CHARACTER\_CREATION (1)
- ☐ 5- HUMAN FACTORS
  - ASSERTIVENESS
  - AWARENESS
  - ARGUMENTATIVE
  - DISTRACTION
    - DISTRACTED
    - DISTRACTION
    - INATTENTION
    - INATTENTIVE
    - UNALERT
    - UNAWARE
    - UNOBSERVANT
    - UNWILGANT
    - UNWATCHFUL
  - MISUNDERSTANDING
  - MISTAKE
  - OMISSION
    - DISREGARDED
    - FORGOT
    - FORGOTTEN
    - IGNORE
    - IGNORED
    - NEGLECTED
    - NEGLIGENCE
    - OMISSION
    - OMIT
    - OMITTED
    - OVERLOOKED
  - POSITIVES
    - BRILLIANT
    - DAZZLING
    - DELIGHTFUL
    - ENABLE
    - EXCELLENT
    - FAST
    - GOOD
    - INTERESTING
    - LIKE\_IT
    - LOVE
    - NICE
    - WORKS\_FLAWLESS
    - WORTHY
  - NEGATIVES
    - ARDUOUS

## Content analysis method

### PROS

- Can measure more accurately
- Can be focused (multi-focus)
- Allows full automation

### CONS

- Dictionary construction & validation

## Text Analytics Challenge

### **THREE MAJOR OBSTACLES**

1) Very large number of word forms

2) Polymorphy of language

One idea → multiple forms

## Text Analytics Challenge

### **THREE MAJOR OBSTACLES**

1) Very large number of word forms

2) Polymorphy of language

One idea → multiple forms

3) Polysemy of words

One word → many ideas



## Challenge #3 – Polysemy of words

### Keyword in Context List (KWIC)

RECNO		KEYWORD	
2766	UND TO THE PASSENGERS IN HOPES OF ALLEVIATING THEIR	STRESS	. THE PART ARRIVED AT THE AIRRAFT AT APPROXIMATELY 1
2690	is still not seated. I had to make a second PA announcement to	stress	the need for all pax to take their seats at this time without furth
209	prior to seatbelt sign going on. I reiterated the announcemnt to	stress	that everyone one would not be able to get up later due to the
3068	is signed off. And we left in a matter of minutes. I would like to	stress	the OUTSTANDING job of everyone on the ground in FLL, espe
7922	recovering from brain surgery and wasn't suppose to be under	stress	and wasn't going to discuss it any further. We were able to pl
8702	and eyes rolled back. F4 - Passenger was had been under	stress	, tired and had very little food. She had taken dramamine and I
7972	d very tired, 3 days with no good sleep. Death in family under	stress	. No allergies, no medication on board, gave him oxygen, we r
8702	F2 - Passenger was under	stress	, tired and had little food. She took dramamine and had 1 merlo
387	ea were in agreement that row ten was causing unnecessary	stress	. I have previously worked as a flight attendant and between
3512	er there either. I finally got through and got Maureen again and	stressed	we had everything but the flight plan filed. She said she would
5448	Mr Koch took the fit back to Den even after we explained and	stressed	the fact that although we would accomodate her for the night
1231	her behavior but I also took into consideration that she must be	stressed	out from the holiday. They advised me (the GSC) that there wa
10273	endants have the proper training in this regard, as it should be	stressed	more than it apparently is. I would like this report to be consi

### Senses of word “stress”

- #1 (psychology) a state of mental or emotional strain or suspense
- #2 (physics) force that produces strain on a physical body
- #3 Verb - single out as important

## Challenge #3 – Polysemy of words

### Keyword in Context List (KWIC)

RECNO		KEYWORD	
2766	UND TO THE PASSENGERS IN HOPES OF ALLEVIATING THEIR	STRESS	. THE PART ARRIVED AT THE AIRRAFT AT APPROXIMATELY 1
2690	is still not seated. I had to make a second PA announcement to	stress	the need for all pax to take their seats at this time without furth
209	prior to seatbelt sign going on. I reiterated the announcemnt to	stress	that everyone one would not be able to get up later due to the
3068	is signed off. And we left in a matter of minutes. I would like to	stress	the OUTSTANDING job of everyone on the ground in FLL, espe
7922	recovering from brain surgery and wasn't suppose to be under	stress	and wasn't going to discuss it any further. We were able to pl
8702	and eyes rolled back. F4 - Passenger was had been under	stress	, tired and had very little food. She had taken dramamine and I
7972	d very tired, 3 days with no good sleep. Death in family under	stress	. No allergies, no medication on board, gave him oxygen, we r
8702	F2 - Passenger was under	stress	, tired and had little food. She took dramamine and had 1 merlo
387	ea were in agreement that row ten was causing unnecessary	stress	. I have previously worked as a flight attendant and between
3512	er there either. I finally got through and got Maureen again and	stressed	we had everything but the flight plan filed. She said she would
5448	Mr Koch took the fit back to Den even after we explained and	stressed	the fact that although we would accomodate her for the night
1231	her behavior but I also took into consideration that she must be	stressed	out from the holiday. They advised me (the GSC) that there wa
10273	endants have the proper training in this regard, as it should be	stressed	more than it apparently is. I would like this report to be consi

### Disambiguation using phrases

**STRESS\*\_THE** or **STRESS\*\_THAT** → “single out as important”

**UNDER\_STRESS** → Emotional State

## Challenge #3 – Polysemy of words

### Keyword in Context List (KWIC)

RECNO		KEYWORD	
2786	UND TO THE PASSENGERS IN HOPES OF ALLEVIATING THEIR	STRESS	. THE PART ARRIVED AT THE AIRCRAFT AT APPROXIMATELY 1
2690	is still not seated. I had to make a second PA announcement to	stress	the need for all pax to take their seats at this time without furth
209	prior to seatbelt sign going on. I reiterated the announcemnt to	stress	that everyone one would not be able to get up later due to the
3068	is signed off. And we left in a matter of minutes. I would like to	stress	the OUTSTANDING job of everyone on the ground in FLL, espe
7922	recovering from brain surgery and wasn't suppose to be under	stress	and wasnt going to discuss it any further. We were able to pl
8702	and eyes rolled back. F4 - Passenger was had been under	stress	tired and had very little food. She had taken dramamine and I
7972	. very tired, 3 days with no good sleep. Death in family under	stress	. No allergies, no medication on board, gave him oxygen, we g
8702	F2 - Passenger was under	stress	tired and had little food. She took dramamine and had 1 merlo
387	ea were in agreement that row ten was causing unnecessary	stress	. I have previously worked as a flight attendant and between
3512	er there either. I finally got through and got Maureen again and	stressed	we had everything but the flight plan filed. She said she would
5448	; Mr.Koch took the fit back to Den even after we explained and	stressed	the fact that although we would accomodate her for the night.
1231	her behavior but I also took into consideration that she must be	stressed	out from the holiday. They advised me (the GSC) that there wa
10273	endants have the proper training in this regard, as it should be	stressed	more than it apparently is. I would like this report to be consk

### Disambiguation using rules

TRANSFER\* **IS NEAR** TECHNOLOGY

TRANSFER\* **IS NOT NEAR** BUS

SATISFIED **IS AFTER** #NEGATION

## Challenge #4 – Misspellings

78,159 word forms:

- 46,404 “unknown” words
  - 75 % misspellings (≈ 35,000)
  - 21 % proper names (products & people)
  - 4% acronyms

## Challenge #4 – Misspellings

### 61 ways to be “Enthusiastic”

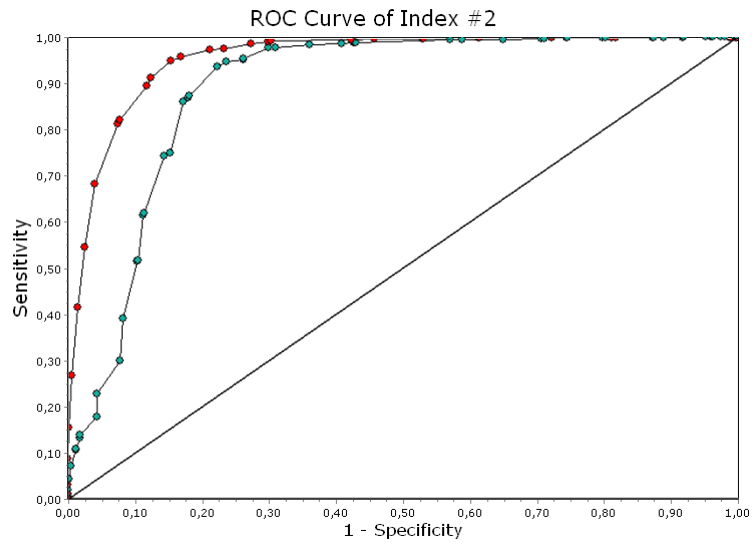
EHNTHUSIASTIC	2	ENTHUSIAIATIC	1	ENTHUSIASITC	1
ENHTUSIASTIC	1	ENTHUSAISTIC	8	ENTHUSICASTIC	1
ENTHEUSIASTIC	1	ENTHUSAITIC	1	ENTHUSICATIC	3
ENTHHUSIATIC	1	ENTHUSASITIC	1	ENTHUSIASIATIC	1
ENTHIASTIC	1	ENTHUSASTIC	52	ENTHUSISATIC	2
ENTHISIASTIC	2	ENTHUSATIC	11	ENTHUSISTIC	7
ENTHOUSIASTIC	13	ENTHUSIACTIC	4	ENTHUSTATIC	2
ENTHSIASTIC	2	ENTHUSIADTOC	3	ENTHUSIASTIC	17
ENTHSUASTIC	1	ENTHUSIAITIC	1	ENTHUSUASTIC	4
ENTHUAISTIC	1	ENTHUSIANSITIC	3	ENTHUTIASTIC	2
ENTHUASASTIC	1	ENTHUSIASITC	9	ENTTHUSIASTIC	1
ENTHUASIASTIC	2	ENTHUSIASITIC	5	ENTUISASTIC	1
ENTHUASIA TIC	2	ENTHUSIASTC	5	ENTUSIASITC	1
ENTHUASISTIC	1	ENTHUSIASTCI	1	ENTUSIASHTIC	2
ENTHUASTIC	30	ENTHUSIASTICE	2	ENTUSIASTIC	47
ENTHUDIASTIC	1	ENTHUSIASTICS	2	ENTUSIATIC	1
ENTHUIASTIC	20	ENTHUSIASTTIC	1	ENUTHUSIASTIC	1
ENTHUISASTIC	2	ENTHUSIAHTIC	1	EUNTHUSIASTIC	1
ENTHUISIASTIC	3	ENTHUSIATIC	185	ANTHUSIASTIC	1
ENTHUSIASTIC	3	ENTHUSIATSIC	4	ENTUSIASIC	1
ENTHUSIASTIC	3			ETHUSIASTIC	28

## Challenge #4 – Misspellings

### Fuzzy and phonetic string comparison algorithms:

- Damerau-Levenshtein
- Koelner Phonetik
- SoundEx
- Metaphone
- Double-Metaphone
- NGram
- Dice
- Jaro-Winkler
- Needleman-Wunch
- Smith-Waterman-Gotoh
- Monge-Elkan

## Challenge #4 – Misspellings



## Challenge #4 – Misspellings

The screenshot shows the 'Suggest 762/896' software interface. The main window displays a list of categories and words under the heading 'CATEGORIES & WORDS'. The list includes various words and their frequencies, such as 'CLARITY' (1), 'CLEAR' (2), 'CLEARER' (1), 'CLEARLY' (1), 'PRECISE' (1), 'COMPREHENSIVE' (1), 'RESPONSIVE' (1), 'WELL DELIVERED' (1), 'NEGATIVE' (1), 'BAD QUALITY' (1), 'ABLE' (1), 'ABOUT' (1), and 'ABOVE' (2). The interface also includes a search bar, a settings menu, and a list of actions to be performed, such as 'Replace in text', 'Add to Substitution', 'Categorize', 'Add to Spell Dictionary', 'Remove Action', and 'Keyword-in-Context'. The status bar at the bottom indicates '89337 cases' and 'Click the search button to find phrases'.

