

# Deriving web genres from text types: a corpus-based approach

*Marina Santini*

University of Brighton (UK)

Presented by

*Davide Mazzi*

Università di Modena e Reggio Emilia (Italy)

AAACL 2006 - Northern Arizona University,

Flagstaff (AZ, USA)

October 20 -22, 2006

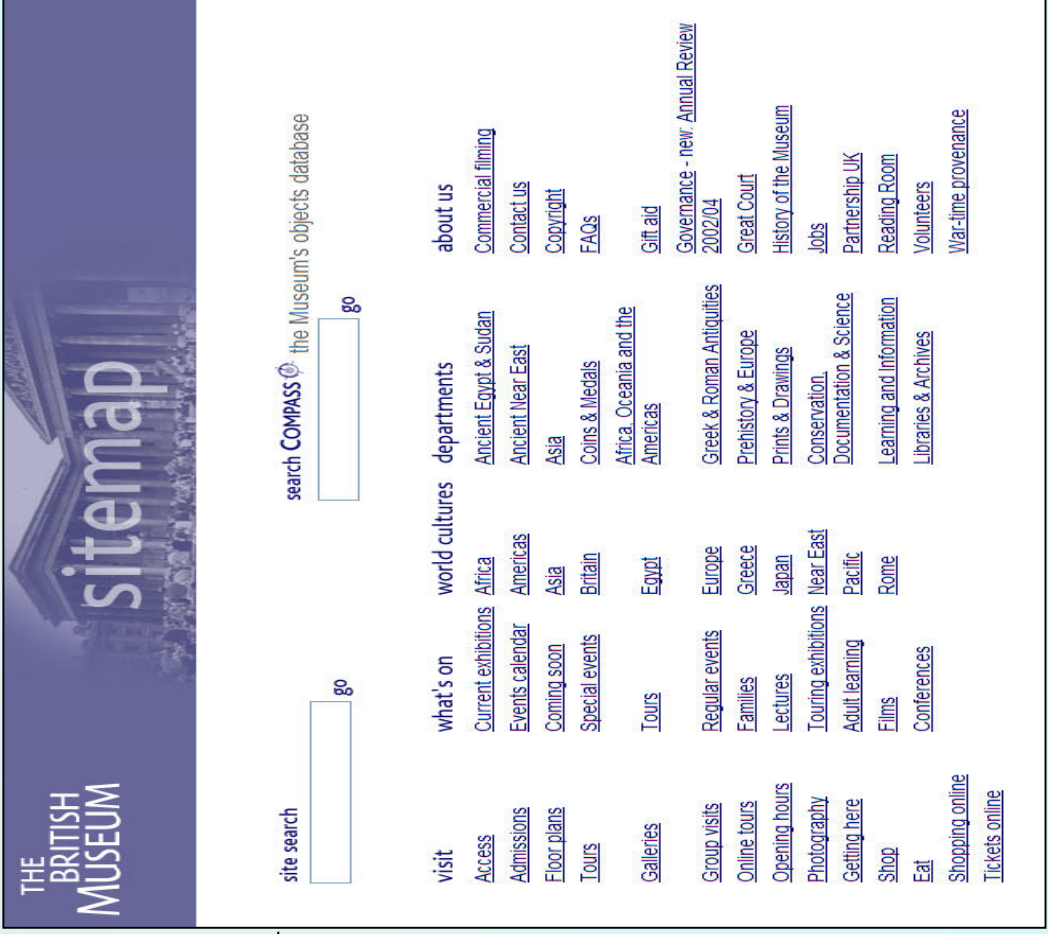
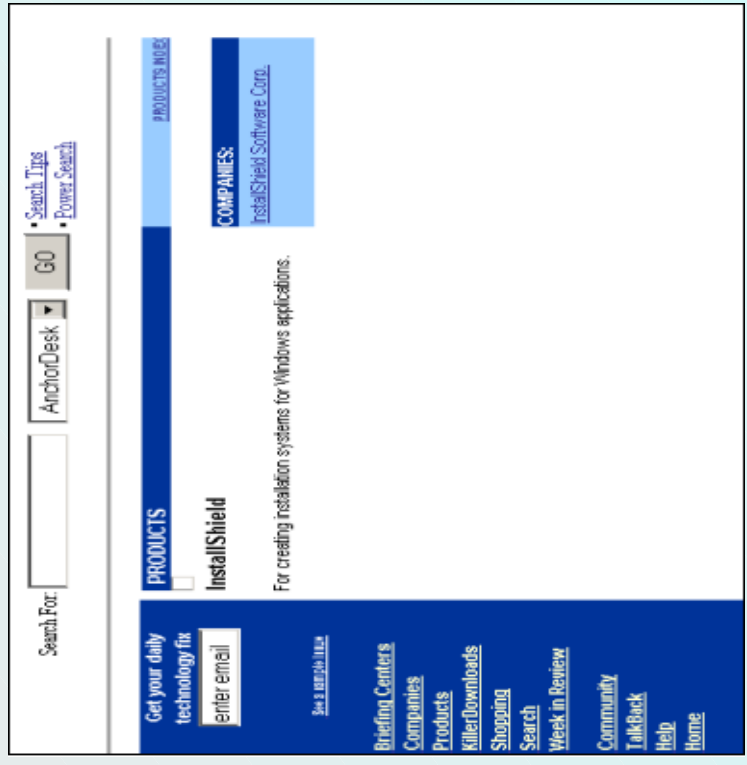
# Benefits

- Parsing accuracy
- Tagging accuracy
- Word sense disambiguation
- IR, IE, Digital Libraries, etc.



# Zero-to-Multi-Genre Classification Scheme

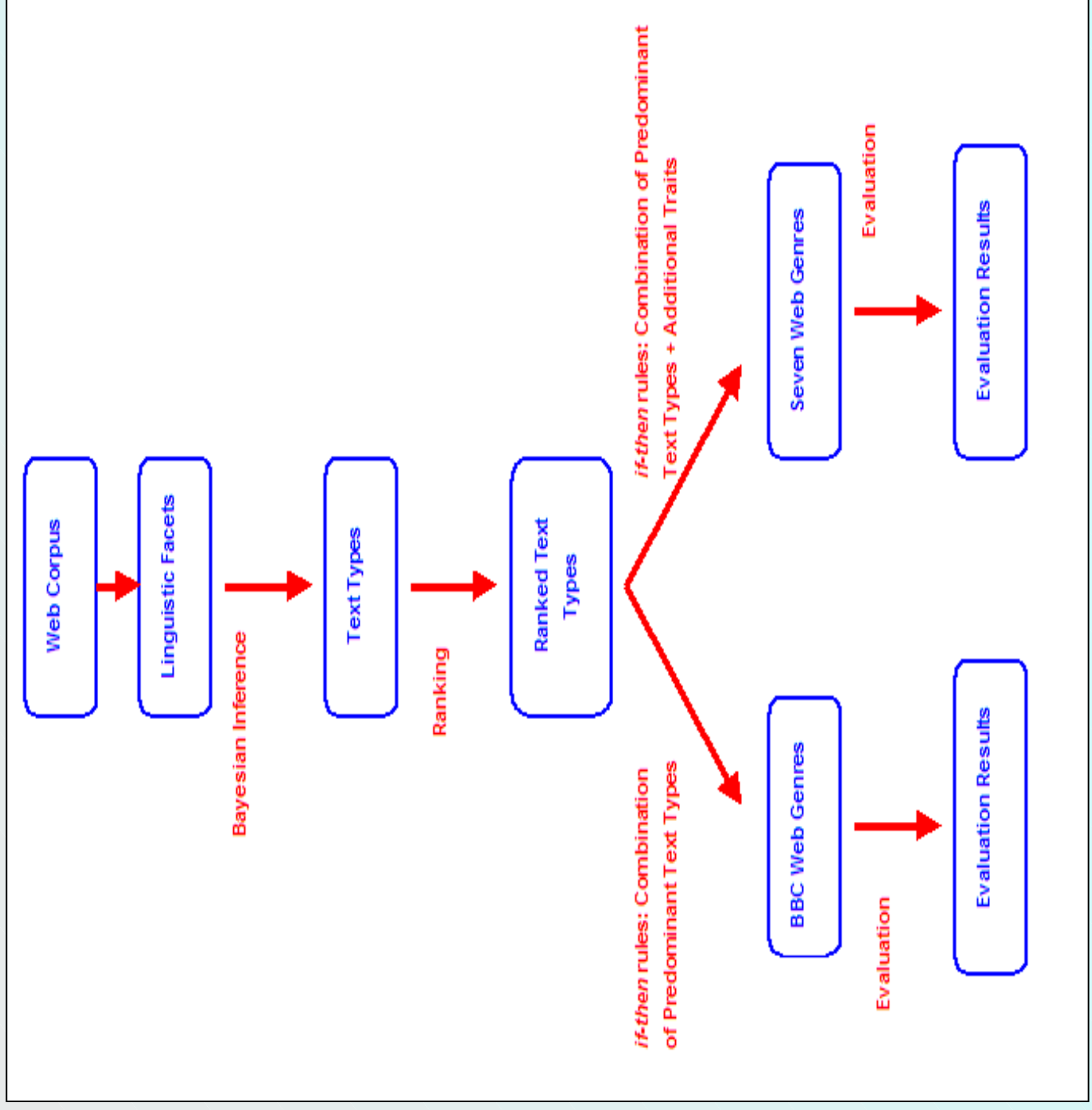
- Zero-genre classification
- Single-genre classification
- Multi-genre classification



# Text Types and Genres

- **Text Types:** rhetorical/discourse patterns,  
e.g. *narration, instruction, argumentation, etc.*
- **Genres:** text categories, e.g. *academic paper, editorial, reportage, blog, FAQs, personal home page, etc.*

# P i p e l i n e



# Web Corpus

Genres	Number of Web Pages	Proportions
Random web pages from the SPIRIT collection	1000	40.32%
Blogs	200	8.065%
Eshops	200	8.065%
FAQs	200	8.065%
Front pages	200	8.065%
Listings	200	8.065%

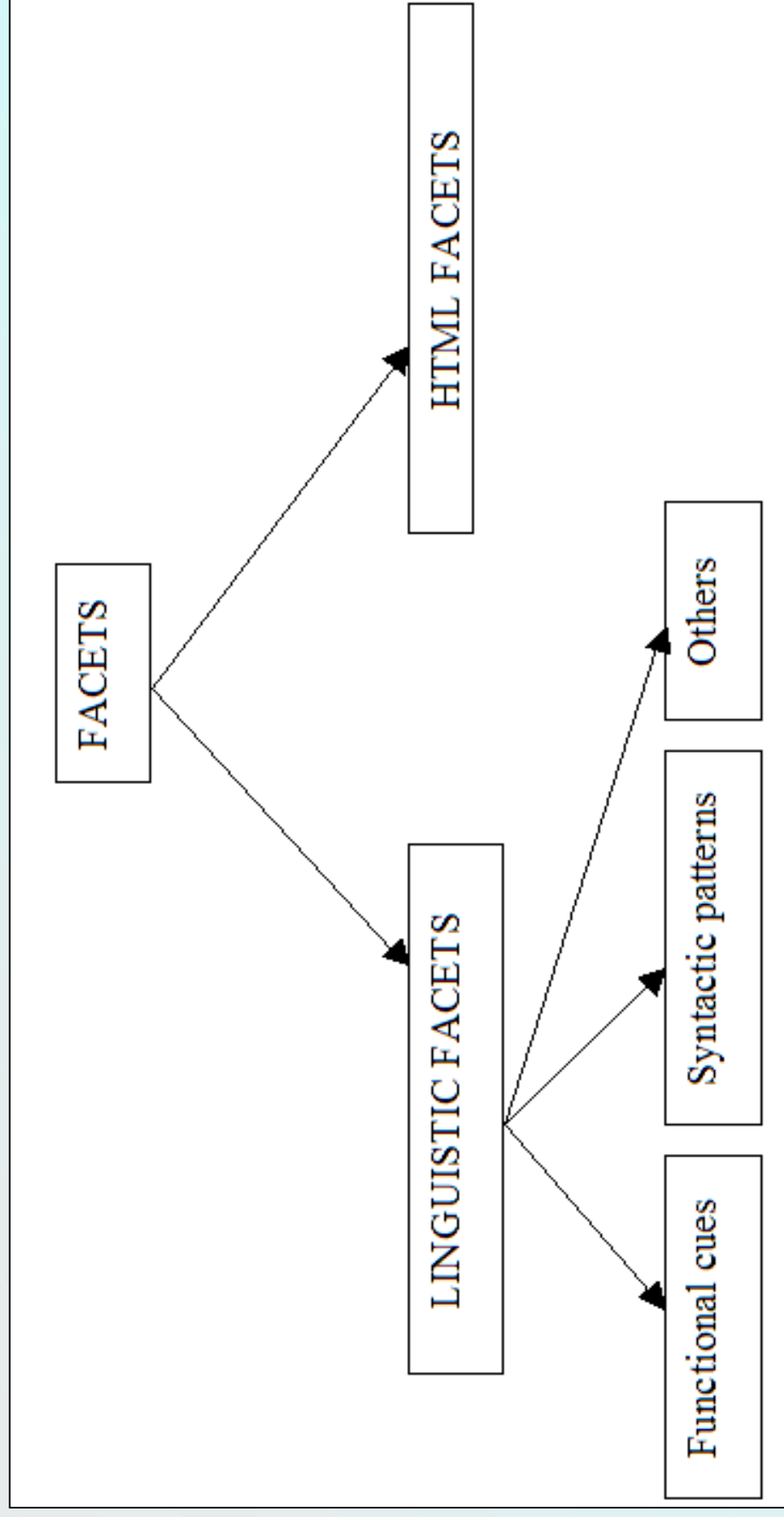
Personal Home Pages	200	8.065%
Search Pages	200	8.065%
BBC Editorials	20	0.806%
BBC DIY mini-guides	20	0.806%
BBC Short Biographies	20	0.806%
BBC Features	20	0.806%
<b>Total</b>	<b>2480</b>	<b>100%</b>

# What is a Facet?

Facets are inspired to Biber's features.

*“The notion of function is closely associated with the notion of situation. A primary motivation for analysis of the components of situation is the desire to link the functions of particular linguistic features to variation in the communicative situation” (Biber, 1988: 33).*

# Types of Facets



# Examples of Facets

## First Person Facet

```
PRON PERS SG1 # Personal pronouns, singular
# I, me, myself, my, mine.

PRON PERS PL1 # Personal pronouns, plural
# we, us, ourselves, our, ours.
```

## Blog words:

```
web log
weblog
blog
journal
diary
posted by
comments
archive

-----

jan
january
feb
february
mar
march
apr
april
may
jun
june
jul
july
aug
august

-----

sept
september
oct
october
nov
november
dec
december

-----

mon
monday
tues
tuesday
wed
weds
wednesday
thurs
thursday
fri
friday
sat
saturday
sun
sunday
```

## CONCESSIVE CLAUSE, INITIAL POSITION

```
although@CS.*?SUBJ.*?FMAINV.*?$$ v_main
```

Ex: Although he had just joined the company, he was treated exactly like all the other employees

```
though@CS.*?@NH , .*?$$ v_main
```

Ex: Though well over eighty, he can walk faster than I can.

```
even though@CS.*?SUBJ.*?FMAINV.*?$$ v_main
```

Ex: Even though you dislike ancient monuments, Warwick Castle is worth a visit.

```
even though@CS.*?@FMAINV_EN.*?$$ v_main
```

Ex: Even though given every opportunity, they would not cooperate with us.

```
even if@CS.*?SUBJ.*?FMAINV.*?$$ v_main
```

Ex: Even if you dislike ancient monuments, Warwick Castle is worth a visit.

```
whereas@CS.*?SUBJ.*?FMAINV.*? , .*?$$ v_main
```

Ex: Whereas the amendment is enthusiastically supported by a large majority in the Senate, its fate is doubtful in the House.

## FUNCTIONALITY FACET

```
<applet # calls a java applet
<bgsound # background sound (obsolete)
<button # create a button in a form
<embed # embed multimedia (obsolete)
<fieldset # groups together fields in a form
<form # create a form for user input
<input # an input element in a form
<legend # provide a legend
<noscript # for browsers not support scripts
<object # can subsume images, applets etc.
<option # a selectable option in a form
<optgroup # provide a hierarchy of choices
<param # a parameter passed to an object
<script # insert an inline script
<select # option selector element in a form
<textarea # a freeform text entry in a form
<var # a program variable
mailto # link to an email address
```

# The Approach

- Inferential model: deductive + inductive.
- The effort is to combine *the classificatory and the descriptive frameworks* (cf. Biber, 1994: 37).

# Methodology (I)

- 4 broad text types:
  - *descriptive\_narrative*
  - *expository\_informational*
  - *argumentative\_persuasive*
  - *instructional*

# Methodology (II)

- *if-then* rules
- 4 BBC web genres: *editorials, Do-It-Yourself (DIY) mini-guides, short biographies, and feature articles*
- 7 novel web genres: *blogs, eshops, FAQs, front pages, listings, personal home pages, and search pages*

# Two Hypotheses

1. The combination of two predominant text types, i.e. the top-ranked text types, is sufficient to derive BBC web genres, more regular in their textuality.
2. The combination of two predominant text types, i.e. the top-ranked text types, plus a combination of additional traits is sufficient to derive novel web genres, which show a textuality more influenced by the interaction allowed by the web.

# Inferring with Odds-Likelihood: Steps

1. Feature extraction and normalization.
2. Conversion of the normalized frequencies into z-scores.
3. Conversion of z-scores into probabilities.
4. Calculation of prior odds from prior probabilities of a text  
type:  $PrOdds(H) = PrProb(H) / 1 - PrProb(H)$
5. Calculation of **multipliers** for the features.
6. Calculation of posterior odds:  
 $Odds(H) = PrOdds(H) * M(E_1) * M(E_n)$
7. Calculation of the probability from odds:  
 $Prob(H) = Odds(H) / 1 + Odds(H)$

# Inferring with Odds-Likelihood:

## Multipliers

if  $\text{Prob}(E) \geq 0.5$  then:

$$M(E) = 1 + (\text{LS} - 1) (\text{Prob}(E) - 0.5) / 0.25$$

if  $\text{Prob}(E) < 0.5$  then:

$$M(E) = 1 - (1 - \text{LN}) (0.5 - \text{Prob}(E)) / 0.25$$

# Gradations of Text Types

1	file_name	descriptive_narrative_expository_informative_persuasive_ranking	0.74	rk_high_instructonal_1-rk_low_argumentative_persuasive_2
2	BBC_DIY_guide_0001	0.17	0.37	0.36
3	BBC_DIY_guide_0002	0.15	0.22	0.36
4	BBC_DIY_guide_0003	0.15	0.22	0.36
5	BBC_DIY_guide_0004	0.26	0.20	
6	BBC_DIY_guide_0005	0.21	0.25	
7	BBC_DIY_guide_0006	0.17	0.16	
8	BBC_DIY_guide_0007	0.09	0.16	
9	BBC_DIY_guide_0008	0.24	0.29	
10	BBC_DIY_guide_0009	0.21	0.32	
11	BBC_DIY_guide_0010	0.16	0.32	
12	BBC_DIY_guide_0011	0.27	0.15	
13	BBC_DIY_guide_0012	0.10	0.13	
14	BBC_DIY_guide_0013	0.13	0.20	
15	BBC_DIY_guide_0014	0.13	0.15	
16	BBC_DIY_guide_0015	0.14	0.21	
17	BBC_DIY_guide_0016	0.10	0.18	
18	BBC_DIY_guide_0017	0.10	0.11	
19	BBC_DIY_guide_0018	0.32	0.20	
20	BBC_DIY_guide_0019	0.14	0.14	
21	BBC_DIY_guide_0020	0.18	0.23	

bbc.co.uk **Homes** [TV and radio](#) [A to Z index](#) [Talk](#) [Search](#)

**DIY guide**

**Bleeding a Radiator**

**Project:** Plumbing  
**Skill level:** Advanced  
**Duration:** 1 day  
**Programme:** BBCi Homes  
**Expert:** Federation of Master Builders

**Tools you will need:**  
 brass radiator key  
 kitchen roll or an old teacloth or towel

**Radiator key**  
 Buy a brass radiator key (this is much stronger than a cheap type). You will also need kitchen roll or an old teacloth or towel. These items will get very stained and you will probably need to throw them away.

**The bleed screw**  
 Find the bleed screw on the radiator. This is a small four sided screw head set 4mm-5mm approximately into a threaded nut in the radiator.

**Check the back**  
 It can be found on any of the sides of the radiator, but if it is on the back, a different key will be required. This looks more like a small spanner with a four sided hole in it.

**Open the screw**  
 The screw should open by turning it anti-clockwise. Unscrew it very slowly to avoid complete removal. This is of utmost importance as the screw is very small, and when it is removed it will allow water out of the radiator at great speed.

[BBC Homepage](#)  
[Lifestyle](#)  
[Homes](#)  
[DIY](#)  
[DIY guide](#)  
[TV and radio](#)  
[Talk](#)  
[Newsletter](#)  
[Contact Us](#)  
[Like this page?](#)  
[Send it to a friend!](#)



# *if-then* Rules

```
if (text_type_1=instructional_1 | argumentative_persuasive_1)  
if (text_type_2=instructional_2 | argumentative_persuasive_2)
```

```
then good DIY candidate
```

```
# -----
```

```
if (text_type_1=descr_narrat_1 | argum_pers_1)  
if (text_type_2=descr_narrat_2 | argum_pers_2)  
if (page_length=LONG)  
if (blog_words >= 0.5 probabilities)
```

```
then good blog candidate
```

# Evaluation of BBC Web Genres

- DIYs → 95%
- Short biographies → 85%
- Editorials → 75%
- Features → 60%

# Evaluation of 7 Novel Web Genres

Web genres	SVM Classifier	Naïve Bayes Classifier	Inferential Model
Blogs	96%	92%	91%
Eshops	88%	76%	83%
FAQs	94.5%	67%	88.5%
Front Pages	100%	98%	97%
Listings	80%	29%	75.5%
Pers. Home Pages	79%	27%	77%
Search Pages	85%	82%	88%
<b>TOTAL</b>	<b>ca. 89%</b>	<b>ca. 67%</b>	<b>ca. 86%</b>

# A snippet from the actual output...

file_name	descriptive	expository	argumenta	instructional
blog_br_107	0.87	0.49	0.87	0.77
suggested_webgenres				
GOOD blog	BAD esho	GOOD faq	BAD front	BAD listing
			BAD php	cBAD spage candidate

Thursday, February 03, 2005

## How to Get Your Way 100% of the Time

WARNING..... WARNING..... WARNING.....

Before continuing with this reader-restricted post, you must verify the following:

- a) I am reasonably competent, at least equally as competent as a 10 year old dog.
- b) I am married or otherwise in a contractual type relationship that is eating away at my insides every single day of my life and as cruel fate would have it, there seems to be no mercy in the foreseeable future.
- c) There are times when homicide seems like a viable option when dealing with the other person in this relationship it gets so fucking bad that 7 years to life in the big house doesn't seem like that big a deal.
- d) The photo below could be you anytime you are dealing with the other person.

ot this shit again!



Sometimes you really don't want to get inside another person's head. This is the perfect example.

## About Me

Name: Bitch U. Crazy  
Location: FAKEville, WTF You Know Where I live, Mozambique

Where have all the same people gone? I really hope they stay there.

[View my complete profile](#)

## Previous Posts

[How to Get Your Way 100% of the Time](#)

[Hicktown Tourette's](#)

[The Joke's on YOU Type A](#)

[Just Get Conceived, We'll Do the Rest!](#)

[Ms. Hushmoney Don't Do Christmas](#)

[Demon Lamp Determined to Burn](#)

[Down the House](#)

[Bizarre Domestic Case: Man Gets](#)

[Pummelled After Pushing Wife "Just a Tad Too Far"](#)

[The Separation of Church and Sanity](#)

[Customer Service Has Left the Building](#)

[Got News! YOUR Kid Isn't THE](#)

[Walking Specimen of the Century, -&#@!ng Jerk Dog!](#)

## Archives

[September 2004](#)

[October 2004](#)

[November 2004](#)

# Conclusions

This inferential model is a starting point for a *zero-to-multi-genre classification scheme*.

Thank you for your attention...