

Table of contents

Introduction	9
1. Background	10
1.1. Previous ESS research initiatives on web scraping of business information	10
1.2. Legal background: the Drones Regulation in Europe	12
1.3. Analysing the industry through company websites	12
1.4. Exploratory analyses of the drone sector	15
2. Approaches to searching for company websites	18
2.1. Searching for websites that provide an overview of drone companies	19
2.2. Searching for websites of individual drone companies	19
2.3. Finding drone websites via pictures of drones	19
2.4. Finding drone websites via a company's impressum page	20
2.5. Finding drone websites using a non-search engine approach	20
3. URL retrieval	22
3.1. URL composition explained	22
3.2. URL collection methodology	23
3.2.1. Search queries	23
3.2.2. Search engines	24
3.2.3. Sequence of steps	27
3.3. Results	31
4. Classification of drone company websites	34
4.1. Drone website classification approaches	34
Classification approaches considered	35
4.1.1. Word-based approach	35
4.1.2. PUlearning approach	36
4.1.3. Supervised Machine Learning approach	37
4.2. Generalising the model	39
4.2.1. The Spanish model	39
4.2.2. Creating a (more) generic model	40
4.2.2.1. Applying the model to Irish websites	40
4.2.2.2. Applying the model to Italian websites	42
5. Data extraction	46
5.1. Information extracted from potential Drone websites	46
5.1.1. Name of the website owner	47
5.1.2. Short description of the website	47

5.1.3. Contact address.....	47
5.1.4. Region of location	48
5.1.5. Email address	48
5.1.6. Telephone numbers.....	49
5.1.7. VAT number.....	49
5.1.8. Activities reported	49
5.1.9. Social media presence	50
5.1.10. E-commerce activity.....	50
5.1.11. Job advertisement presence	50
5.1.12. website Start date	50
5.1.13. Is the website really about drones?	50
5.1.14. Is the website's owner active in the country studied?	51
5.1.15. Is it a company website?.....	51
5.1.16. Is it the website of a drone company active in the country?	52
5.1.17. Comparison of findings	52
5.2. Results.....	53
6. Generalising the methodology and tools	59
6.1. Main considerations.....	59
6.2. Generic keyword extraction	60
6.2.1. Method.....	60
6.2.2. Results.....	61
6.3. Discussion	65
7. Conclusions and recommendations	66
References:	69
Annex I: Creating a list of drone company websites for Spain.....	72
Annex II: Hard- and software requirements	73
Annex III: Top 20 features in the Logistic Regression model	75
Annex IV: URLs of companies found for Spain	76
Annex V: URLs of companies found for Italy.....	78
Annex VI: URLs of companies found for Ireland	80
Annex VII: Generic keyword extraction: implementing the method	81