

Εξόρυξη κειμένων με τη χρήση του λογισμικού Orange

Παναγιώτης Πράσιнос

ΠΕΡΙΛΗΨΗ

Η παρούσα διπλωματική εργασία πραγματεύεται τις εφαρμογές της μηχανικής μάθησης για την αναγνώριση και κατηγοριοποίηση κειμένων. Οι εξελίξεις στον τομέα της μηχανικής μάθησης και γενικά στον τομέα της τεχνητής νοημοσύνης και της επιστήμης δεδομένων είναι σημαντικές και επιτρέπουν μεταξύ άλλων την αυτόματη επεξεργασία μεγάλου όγκου δεδομένων, είτε αυτά αφορούν σε εικόνες, βίντεο, ήχο είτε αφορούν σε κείμενα και την εξαγωγή πολύτιμων πληροφοριών από αυτά. Η μηχανική μάθηση επιτρέπει την αναγνώριση μοτίβων, την αναγνώριση του περιεχομένου και την κατηγοριοποίηση των δεδομένων με τέτοιο τρόπο που μπορεί να παρουσιάζει μεγάλο ενδιαφέρον για τον ερευνητή. Έτσι η επεξεργασία δεδομένων γίνεται πολύ ταχύτερα, κάτι το οποίο μπορεί να ωφελήσει τόσο ερευνητικούς σκοπούς όσο και σκοπούς σε διάφορους τομείς της ζωής όπως η μηχανική, η οικονομία, η ιατρική, η ασφάλεια κ.α.

Στην παρούσα διπλωματική εργασία πραγματοποιήθηκε ενδελεχής ανασκόπηση της βιβλιογραφίας όσον αφορά την μηχανική μάθηση και την εξόρυξη δεδομένων από κείμενα. Μετά εστίασαμε στην ανάλυση πιο συγκεκριμένων κατηγοριών. Πρόκειται για κάποιες τεχνικές οι οποίες είναι ιδιαίτερες χρήσιμες σε ένα ευρύ φάσμα διαδικασιών που αφορούν την ταξινόμηση, την επεξεργασία και γενικά την άντληση πληροφοριών από κείμενα.

Οι τεχνικές που προαναφέρθηκαν μπορούν να εφαρμοστούν με πολλούς και διαφορετικούς τρόπους. Υπάρχουν αρκετά λογισμικά τα οποία μπορούν με ποικίλους τρόπους να εμφανίσουν εξίσου καλά αποτελέσματα, αντίστοιχα με αυτά που παίρνουμε με τη δημιουργία κώδικα γραμμένου σε μία γλώσσα προγραμματισμού όπως python, c++ κ.τ.λ.

Τμήμα Μηχανολόγων και Αεροναυπηγών Μηχανικών – Τομέας Διοίκησης & Οργάνωσης

Ένα από τα λογισμικά που δεν απαιτούν την ιδιαίτερη γνώση κάποιας γλώσσας προγραμματισμού είναι το Orange. Είναι μία ανοιχτή πηγή οπτικοποίησης δεδομένων και χρησιμοποιήθηκε σε αυτή τη διπλωματική για να εφαρμοστούν στην πράξη οι τεχνικές που ειπώθηκαν πιο πάνω. Δημιουργήσαμε ένα μοντέλο για κάθε περίπτωση που εξετάζουμε, το οποίο μπορεί να δεχθεί ως είσοδο κείμενα σε διάφορες μορφές. Αυτά, όμως, θα τα δούμε αναλυτικά στη συνέχεια. Σε αυτό το σημείο πρέπει να πούμε ότι γίνεται αναφορά και σε άλλα λογισμικά, παρόμοια με το Orange, τα οποία μπορούν να επιδείξουν εξίσου καλή εφαρμογή σε παρόμοιες διαδικασίες.

Λέξεις κλειδιά

μηχανική μάθηση, επεξεργασία κειμένων, εξόρυξη δεδομένων, κατηγοριοποίηση κειμένων, Orange

Text mining using the Orange tool

Panagiotis Prasinos

ABSTRACT

This thesis deals with the applications of machine learning for text recognition and text classification. The development in the field of machine learning and in the field of artificial intelligence and data science is significant and allow, among other things, the automatic processing of large volumes of data, whether they relate to images, videos, audio or texts and extract some valuable information from them. Machine learning allows the recognition of patterns, the recognition of content and the classification of data in such a way that may be of great interest to the researcher. So , data processing is done much faster, something that can benefit research purposes and purposes in various areas of life such as engineering, economics, medicine, security, etc.

In this work we carried out a thorough review of the literature on machine learning and data mining .Then, we focused on the analysis of more specific categories. These are some techniques that are particularly useful in a wide range of procedures related to classification, processing and generally extracting information from texts.

The techniques that mentioned above can be applied in many different ways. There are several softwares that can show good results in a variety of ways, corresponding to what we get by creating code written in a programming language such as python, c ++, e.t.c.

One of the software that does not require special knowledge of a programming language is Orange . It is an open source of data visualization and was used in this thesis to do the practical application of the techniques that mentioned above. We have created a model for each case we were considering , which can accept texts in several forms . We will see all these

in detail later to this paper . At this moment we must say that we made a reference to other software , similar to Orange , which can show just as good application to similar processes .

Λέξεις κλειδιά

machine learning , text processing , text mining , text classification , Orange