Classificazione di sottogeneri musicali tramite tecniche di text-mining

Lorenzo Mammana 807391 Gabriele Molteni 800900 Matteo Scarpone 800900

***Sommario*—This document is a model and instructions for LaTeX. This and the IEEEtran.cls file define the components of your paper [title, text, heads, etc.]. *CRITICAL: Do Not Use Symbols, Special Characters, Footnotes, or Math in Paper Title or Abstract.**

***Index Terms*—component, formatting, style, styling, insert**

I. Introduzione

Numerosi studi in letteratura sono stati effettuati riguardo la classificazione automatica di generi musicali sfruttando le liriche delle canzoni. Ci si aspetta che generi musicali diversi contengano tematiche e linguaggi molto differenti tra di loro e quindi la classificazione dovrebbe essere possibile. In questo studio si è deciso di verificare la possibilità di estendere questa analisi andando a valutare la possibilità di distinguere sottogeneri musicali di uno stesso genere. In particolare ci si è concentrati sull'analisi di sottogeneri della musica metal, questo tipo di musica presenta una enorme varietà di sottogeneri musicali, ciascuno con il proprio stile linguistico e le proprie tematiche. La classificazione delle liriche è stata effettuata utilizzando tecniche di text-preprocessing e text-mining. Uno dei metodi più noti per la classificazione è quello basato su Latent Dirichlet Allocation (LDA), questo tipo di metodo permette di eseguire l'inferenza di *Topic* all'interno di un testo in maniera non supervisionata. Avendo a disposizione dati etichettati con uno o più sottogeneri si è utilizzato una versione più moderna di LDA, denomita LLDA (Labelled LDA) che permette di etichettare i *topic* ottenendo una classificazione supervisionata.

II. Dataset

Il dataset è stato costruito eseguendo lo scraping di due diversi siti web Darklyrics e Metal Archives, il primo sito è una collezione di liriche, mentre il secondo è stato utilizzato per integrare le informazioni riguardo al genere degli artisti. La prima versione del dataset prodotta è composta da 259166 liriche, ogni lirica è composta dai seguenti campi:

- band - Il nome del gruppo compositore
- album - L'album in cui è contenuta
- year - L'anno di uscita
- song - Il nome della canzone
- lyrics - Le liriche della canzone
- genre - Il genere o i generi della canzone
- lang - La lingua della canzone

Il dataset estratto contiene numerosi campi mancanti come mostrato in tabella I:

Tabella I
Attributi mancanti e unici

Attributo	Mancanti	Valori unici
band	0	8456
album	11	26089
year	46	73
song	784	194215
lyrics	17891	226154
genre	55929	1694
lang	0	37

Tutte le righe con attributi mancanti sono state eliminate dal dataset utilizzato per la classificazione finale. La lingua della lirica viene prodotta con un algoritmo in grado di calcolare la percentuale di appartenza di un testo ad una certa lingua, in tabella I il numero di *lang* mancanti è nullo solamente perchè viene utilizzato il valore MISSING come placeholder. In tabella II viene invece mostrata la distribuzione delle prime dieci lingue contenute nel dataset.

Tabella II
Top 10 frequenza lingue

attributo	frequenza
en	0.796778
MISSING	0.069033
ro	0.041854
de	0.025347
es	0.019949
fi	0.007833
pt	0.005483
fr	0.005336
sv	0.004460
no	0.004121

Per la classificazione finale sono state mantenute esclusivamente le liriche in lingua inglese, le altre lingue sono troppo poco frequenti per poter essere utilizzate e sarebbero un fattore confondente troppo grande per gli algoritmi di classificazione. La rimozione delle liriche mancanti o non in inglese riduce il numero di righe a 200556.

A. Gestione del genere musicale

Come si vede in tabella I il numero di generi univoci presenti all'interno del dataset è estremamente elevato, questo è dovuto al fatto che i dati estratti tramite scraping non sono ben categorizzati, ma possono presentare leggere variazioni o ordinamenti differenti. Per poter rendere fattibile la classificazione è stata utilizzata la conoscenza del dominio per raggruppare i sottogeneri in macro-sottogeneri, in particolare la scelta è stata quella di fissare nove generi principali a cui vengono ricondotte tutte le liriche nle dataset. Questo introduce ovviamente un bias in quanto la categorizzazione è basata esclusivamente su conoscenza a priori riguardante il genere, senza essere influenzata dall'effettivo contenuto delle liriche. Un secondo bias è relativo al fatto che i generi sono associato al gruppo musicale e non al singolo album, ciò che accade nella realtà è che un artista produce album di diversi generi, mentre invece nel dataset questa distinzione non è evidente.

I nove generi proposti sono i seguenti:

- Rock
- Heavy
- Thrash
- Power
- Folk
- Progressive
- Death
- Doom
- Black

I generi sono stati selezionati in base alle pre-conoscenze degli autori in modo tale che possa essere stabilito un ordine tra di essi; nella lista precedente ogni genere è teoricamente contenuto all'interno dell'insieme composto dal genere successivo. Vengono prodotti due dataset utilizzati per il task di classificazione, il primo è un dataset multi-label, mentre il secondo utilizza l'ordinamento per associare una singola label come attributo di output. I dataset prodotti sono fortemente sbilanciati, il dataset multilabel può teoricamente rappresentare $2^9 - 1 = 511$ generi diversi, nel dataset sono presenti solamente 110 di queste possibili combinazioni. Come mostrato in tabella III sono solamente 18 i generi con frequenza superiore all'1%, ci si aspetta quindi che la classificazione su questo dataset sarà particolarmente complicata.

Tabella III
GENERI CON FREQUENZA SUPERIORE ALL'1%

Generi	Frequenze
{Death Metal}	0.214409
{Black Metal}	0.097342
{Power Metal}	0.072505
{Thrash Metal}	0.063777
{Heavy Metal}	0.058732
{Death Metal, Thrash Metal}	0.050035
{Doom Metal}	0.046568
{Death Metal, Doom Metal}	0.037722
{Progressive Metal}	0.031522
{Heavy Metal, Power Metal}	0.031199
{Death Metal, Black Metal}	0.027881
{Heavy Metal, Rock}	0.024937
{Thrash Metal, Power Metal}	0.019998
{Heavy Metal, Thrash Metal}	0.019103
{Progressive Metal, Power Metal}	0.014146
{Doom Metal, Black Metal}	0.012251
{Doom Metal, Heavy Metal}	0.010021

Nel caso singola-label il dataset rimane piuttosto sbilanciato, ma la classificazione dovrebbe essere nettamente più semplice da eseguire. Le frequenze sono mostrate in tabella IV.

Tabella IV
FREQUENZE GENERI SINGOLA LABEL

Genere	Frequenza
Death Metal	47379
Black Metal	28444
Power Metal	21019
Doom Metal	19413
Thrash Metal	13341
Rock	11101
Heavy Metal	9454
Progressive Metal	8757
Folk Metal	2060

É interessante notare l'andamento dei generi nel tempo, il dataset fornisce una buona copertura dei generi per anno, in particolare si vede, come è accaduto realmente, un aumento della "pesantezza" del genere musicale con il passare degli anni (Figura 1).

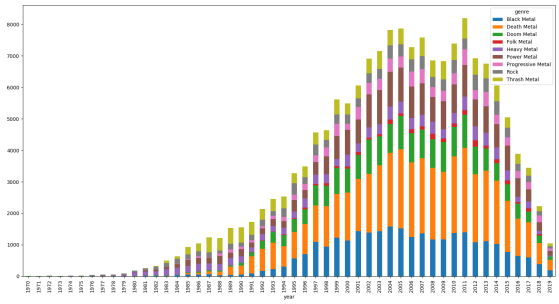

Figura 1. Distribuzione dei generi musicali negli anni

Si è proceduto verificando quali sono lo parole più frequenti per ogni genere, si nota che le dieci parole più frequenti sono piuttosto simili all'interno di ogni genere come visibile in tabella V.

Tabella V
TOP 10 PAROLE FREQUENTI PER GENERE

Rock	Heavy	Thrash	Power	Folk	Progressive	Death	Doom	Black
away	got	death	away	blood	away	blood	away	black
got	know	know	know	come	eyes	death	eyes	blood
know	life	life	life	land	know	eyes	know	death
life	like	like	never	life	life	life	life	eyes
like	never	never	night	like	like	like	like	life
love	night	one	one	never	never	never	never	night
never	one	see	see	night	one	one	one	one
one	see	time	time	one	see	see	see	see
see	take	way	way	see	time	time	time	time
time	time	world	world	time	world	world	world	world

B. Maintaining the Integrity of the Specifications

The IEEEtran class file is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

III. PREPARE YOUR PAPER BEFORE STYLING

Before you begin to format your paper, first write and save the content as a separate text file. Complete all content and organizational editing before formatting. Please note sections III-A–III-E below for more information on proofreading, spelling and grammar.

Keep your text and graphic files separate until after the text has been formatted and styled. Do not number text heads—LaTeX will do that for you.

A. Abbreviations and Acronyms

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, ac, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

B. Units

- Use either SI (MKS) or CGS as primary units. (SI units are encouraged.) English units may be used as secondary units (in parentheses). An exception would be the use of English units as identifiers in trade, such as "3.5-inch disk drive".
- Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.
- Do not mix complete spellings and abbreviations of units: "Wb/m^2" or "webers per square meter", not "webers/m^2". Spell out units when they appear in text: ". . . a few henries", not ". . . a few H".
- Use a zero before decimal points: "0.25", not ".25". Use "cm^3", not "cc".)

C. Equations

Number equations consecutively. To make your equations more compact, you may use the solidus (/), the exp function, or appropriate exponents. Italicize Roman symbols for quantities and variables, but not Greek symbols. Use a long dash rather than a hyphen for a minus sign. Punctuate equations with commas or periods when they are part of a sentence, as in:

$$a + b = \gamma \tag{1}$$

Be sure that the symbols in your equation have been defined before or immediately following the equation. Use "(1)", not "Eq. (1)" or "equation (1)", except at the beginning of a sentence: "Equation (1) is . . ."

D. LaTeX-Specific Advice

Please use "soft" (e.g., `\eqref{Eq}`) cross references instead of "hard" references (e.g., `(1)`). That will make it possible to combine sections, add equations, or change the order of figures or citations without having to go through the file line by line.

Please don't use the `{eqnarray}` equation environment. Use `{align}` or `{IEEEeqnarray}` instead. The `{eqnarray}` environment leaves unsightly spaces around relation symbols.

Please note that the `{subequations}` environment in LaTeX will increment the main equation counter even when there are no equation numbers displayed. If you forget that, you might write an article in which the equation numbers skip from (17) to (20), causing the copy editors to wonder if you've discovered a new method of counting.

BibTeX does not work by magic. It doesn't get the bibliographic data from thin air but from .bib files. If you use BibTeX to produce a bibliography you must send the .bib files.

LaTeX can't read your mind. If you assign the same label to a subsubsection and a table, you might find that Table I has been cross referenced as Table IV-B3.

LaTeX does not have precognitive abilities. If you put a `\label` command before the command that updates the counter it's supposed to be using, the label will pick up the last counter to be cross referenced instead. In particular, a `\label` command should not go before the caption of a figure or a table.

Do not use `\nonumber` inside the `{array}` environment. It will not stop equation numbers inside `{array}` (there won't be any anyway) and it might stop a wanted equation number in the surrounding equation.

E. Some Common Mistakes

- The word "data" is plural, not singular.
- The subscript for the permeability of vacuum μ_0, and other common scientific constants, is zero with subscript formatting, not a lowercase letter "o".
- In American English, commas, semicolons, periods, question and exclamation marks are located within quotation marks only when a complete thought or name is cited, such as a title or full quotation. When quotation marks are used, instead of a bold or italic typeface, to highlight a word or phrase, punctuation should appear outside of the quotation marks. A parenthetical phrase or statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.)
- A graph within a graph is an "inset", not an "insert". The word alternatively is preferred to the word "alternately" (unless you really mean something that alternates).
- Do not use the word "essentially" to mean "approximately" or "effectively".
- In your paper title, if the words "that uses" can accurately replace the word "using", capitalize the "u"; if not, keep using lower-cased.
- Be aware of the different meanings of the homophones "affect" and "effect", "complement" and "compliment", "discreet" and "discrete", "principal" and "principle".
- Do not confuse "imply" and "infer".
- The prefix "non" is not a word; it should be joined to the word it modifies, usually without a hyphen.
- There is no period after the "et" in the Latin abbreviation "et al.".
- The abbreviation "i.e." means "that is", and the abbreviation "e.g." means "for example".

An excellent style manual for science writers is [7].

F. Authors and Affiliations

The class file is designed for, but not limited to, six authors. A minimum of one author is required for all conference articles. Author names should be listed starting from left to right and then moving down to the next line. This is the

author sequence that will be used in future citations and by indexing services. Names should not be listed in columns nor group by affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization).

G. Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is "Heading 5". Use "figure caption" for your Figure captions, and "table head" for your table title. Run-in heads, such as "Abstract", will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced.

H. Figures and Tables

Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation "Fig. 2", even at the beginning of a sentence.

Tabella VI
TABLE TYPE STYLES

Table Head	**Table Column Head**		
	Table column subhead	***Subhead***	***Subhead***
copy	More table copy[a]		

[a]Sample of a Table footnote.

Figura 2. Example of a figure caption.

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity "Magnetization", or "Magnetization, M", not just "M". If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write "Magnetization (A/m)" or "Magnetization {A[m(1)]}", not just "A/m". Do not label axes with a ratio of quantities and units. For example, write "Temperature (K)", not "Temperature/K".

ACKNOWLEDGMENT

The preferred spelling of the word "acknowledgment" in America is without an "e" after the "g". Avoid the stilted expression "one of us (R. B. G.) thanks ...". Instead, try "R. B. G. thanks...". Put sponsor acknowledgments in the unnumbered footnote on the first page.

REFERENCES

Please number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use "Ref. [3]" or "reference [3]" except at the beginning of a sentence: "Reference [3] was the first ..."

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors' names; do not use "et al.". Papers that have not been published, even if they have been submitted for publication, should be cited as "unpublished" [4]. Papers that have been accepted for publication should be cited as "in press" [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

RIFERIMENTI BIBLIOGRAFICI

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.
[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
[3] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
[4] K. Elissa, "Title of paper if known," unpublished.
[5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.
[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.